Computational Assessment of Genetic Variation beyond Single Nucleotide Changes

by
Christopher Douville

A dissertation submitted to Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy

Baltimore, Maryland
May, 2017

Abstract

Advances in sequencing technology have greatly reduced the costs incurred in collecting raw sequencing data and researchers now have access to very large datasets of genomic alterations. Computational tools are necessary in order to interpret and discover biologically relevant genetic variation from sequencing data. Current computational tools, however, have overwhelmingly focused on single nucleotide changes. Much less work has been devoted to computational tools to prioritize insertion and deletion variants and chromosomal abnormalities. Insertion/deletion variants (indels) alter protein sequence and length, yet are highly prevalent in healthy populations, presenting a pressing need for bioinformatics classifiers. Chromosomal abnormalities can produce a wide range of genetic disorders including in miscarriages, developmental disorders, and carcinogenesis. While numerous tools have been developed to detect chromosomal abnormalities, these tools have limited utility at lower cell admixtures.

In this dissertation, I focus on the development of computational approaches beyond single nucleotide variants. I introduce a novel computational approach to assess indels variants (Chapters 2-3). I compare this method to existing computational approaches and investigate potential ways to improve indel prediction. Next, I develop a bioinformatics approach entitled WALDO (Within-sample AneupLoidy DiscOvery) specifically designed to detect chromosomal abnormalities as well as microsatellite instability (Chapters 4-6).

Thesis Advisor: Rachel Karchin

Thesis Reader: Bert Vogelstein

Acknowledgements

This work would not have been completed without contributions from numerous people. First, I would like thank my advisor, Rachel Karchin for giving me an opportunity to join her lab. My undergraduate major in chemical engineering had little training in statistics, programming, or genetics so I am forever grateful for her taking a chance on me. I would also like to thank both Rachel Karchin and Bert Vogelstein for their patience, guidance, and countless hours spent working with me on these projects. I have learned so much from both of you about scientific research, intellectual curiosity, and mentorship. I would also like to thank Ken Kinzler and Nick Papadopoulos for joining my thesis committee and providing valuable insights into these projects. I am forever grateful for all the people I have a worked with during my time at Johns Hopkins, especially all Noushin, Hannah, Violetta, Rohit, Collin, Dave, Yun Ching, Ashok, and the members of the Ludwig Cancer Center. I am very grateful to Maria Popoli and the A-Team who sequenced many of the samples used in this work.

Next, I would like to thank my undergraduate mentors and lab mates Dr's Stewart Wang, Grace Su, Englesbe, Davey Lang, and Chris Brede. Some of my fondest undergraduate memories were rushing from the pool at Canham Natatorium still smelling like chlorine into the Burn and Trauma Center Lab. It was my undergraduate research summers that I fell in love with research and decided to apply to Johns Hopkins. Without your flexibility, I never would have been able to juggle my time as a college athlete, chemical engineering coursework and research obligations.

Lastly, I would like to thank my family. They provided tremendous support along my academic journey, but most importantly after years of competing against Nick, it was only natural to extend our sibling rivalry into our research publications and citations.

Table of Contents

List of Tables

List of Figures

Chapter 1: Overview

With the advent of high-throughput sequencing technology, researchers face a bottleneck in terms of the time required to analyze the potential impact on disease etiology of the many genetic variants routinely detected. Computational algorithms can in principle help researchers to prioritize and direct future work by narrowing down the numerous genetic alterations identified in sequencing studies. Current tools have largely focused on single nucleotide changes and much less work on insertion/deletion (indel) variants and chromosomal abnormalities.

1.1 Computational Methods to Prioritize Single Nucleotide Changes

1.1.1 Non-Synonymous Single Nucleotide Variants

Non-synonymous single nucleotide variants are genetic variants that alter the protein sequence. Experimental assessment of protein activity for mutated proteins is very difficult, and is further impeded by the large number of NS-SNVs revealed by exome sequencing studies. This has motivated the development of many statistical and computational methods for evaluating the functional impact of non-synonymous changes on proteins. The methods fall broadly into two categories, those that score mutations on the basis of biological principles (SIFT [1], MutationAssessor [2], MAPP [3], PANTHER [4], among others), and methods that use existing knowledge about the functional effects of mutations in the form a training set for supervised machine learning (PolyPhen2 [5], SNAP [6], SNPs3D [7], MutPred [8], MutationTaster [9], among others [10]). These methods are known to perform well at distinguishing Mendelian disease mutations from common single

nucleotide polymorphisms [9] and usually offer either a numeric score that represents the predicted functional impact of an amino acid substitution, or a probability that the substitution is deleterious to the protein. Mutation scores can be used to substantially reduce the number of candidate disease-causing mutations detected in exome sequencing studies, but additional evidence is still needed to identify the causal mutation

1.1.2 Single Nucleotide Splice Variants

Ample evidence indicates sequence changes at key regions in pre-mRNA may cause aberrant splicing resulting in disease [11; 12], and single nucleotide variants (SNVs) at splice sites are known to contribute to at least 10% of identified inherited diseases [12]. Importantly, SNVs beyond the splice site can also cause disease by disrupting the binding of splicing machinery and occur in both exons and introns. Current estimates suggest as much as 25% of known Mendelian disease exonic SNVs thought to impair protein function through missense and nonsense changes actually disrupt splicing [13; 14]. There is also evidence that SNVs deep into the intron (further than 50 bp from the exon/intron boundary) can even cause aberrant splicing [15]. Numerous computational methods can predict the impact of splice site SNVs but only a few attempt to identify which additional exonic and intronic SNVs are important for splicing [16-19].

Computational prediction of SNVs at exon/intron junctions is based on strong signals of consensus di-nucleotide conservation, whereas prediction outside of the immediate vicinity of exon/intron junctions will require integration of many complex weaker cis-acting signals [20]. One of these weaker signals is the positional distribution of sequence motifs. Methods for hierarchical clustering of motifs have been developed to identify higher-level patterns of organization [14; 21]. Beyond positional distribution, the

complex network of weaker signals depends on conservation [22] and RNA secondary structure [23].

1.2 Insertion and Deletion Variants

Inherited (germline) and non-inherited (somatic) insertion/deletion variants (referred to as indels throughout) in both the coding and noncoding region of the genome play a critical role in human health. The average human exome contains over four hundred naturally occurring germline micro- indels[24]. In-frame indels account for ~50% of these variants, and result from the insertion/deletion of an integer number of codons, and ultimately amino acids. Frameshifts account for the other ~50% of indels, and result from contiguous nucleotide insertions/deletions of a length not divisible by three. This fractional change in the number of codons shifts the translational reading frame, resulting in an entirely new downstream sequence thereby shifting the position at which the first stop codon is encountered. Thus, frameshift indels translate to protein that is very distinct from the native protein, particularly if the indel occurs early in the transcript sequence. Both in-frame and frameshift indels alter protein sequence and length.

Because they drastically alter protein primary structure, but are also highly prevalent in healthy populations, indels present a unique classification challenge. Clearly, the principles governing pathogenicity are not identical to those governing changes in protein function and stability; otherwise, most indels would be pathogenic. The challenge then arises because protein sequence, structure, function and stability are typically considered when assessing a variant of unknown impact on disease liability [25]. These protein-based criteria could lead to a high false positive rate, and therefore low

specificity, because the fraction of indels that appreciably impact health might be overestimated.

Additionally, somatic indels in both the coding and non-coding region play a key role in cancer. Somatic indels that disrupt DNA repeats indicate a sample has impaired or deficient DNA mismatch repair (MMR) that results in microsatellite instability (MSI). Detection of MSI has a drastic impact on cancer treatment and survivability [26]. The traditional way to detect MSI uses a panel of just 5 markers [27]. Next generation sequencing techniques can analyze a much larger number of locations but analyzing repetitive regions can be computationally challenging [28].

The increased utilization of high-throughput genomic sequencing technologies and hopes for their clinical application, coupled with the high prevalence of indels, has led to a demand for bioinformatic tools connected to indel variants.

1.3 Chromosomal Abnormalities (Aneuploidy)

Aneuploidy is an abnormal number of chromosomes resulting from chromosomes not properly splitting during cell division [29] and can produce a wide range of genetic disorders ranging from spontaneous abortions, Down syndrome (Trisomy 21), Patau syndrome, (Trisomy 13), Edwards syndrome (Trisomy 18), Turner syndrome (XO), Klinefelter syndrome (XXY), Triple X syndrome [29] as well as carcinogenesis [30-32]. There is a pressing need for prenatal and cancer screening computational techniques to reliably detect micro-deletions, duplications as well as whole chromosome gains and losses.

Many aneuploidy detection protocols are invasive procedures such as amniocentesis, chorionic villus, and biopsy. Invasive procedures have a non-negligible risk of fetal loss [33] and can cause severe pain and complications [34]. Genetic screening

using cell free DNA has many advantages over invasive techniques. Circulating cell free DNA can be retrieved from blood, urine, and stool [35]and can include fetal DNA (cffDNA) [36] or tumor DNA (ctDNA) [37]. Genetic screening from fetal or tumor cell free DNA, however, can be challenging due to extremely low admixtures rates from the source of interest. Fetal DNA fraction typically 12% depending on the stage of pregnancy [38] [39; 40] and tumor DNA fraction can range from less to .01% to well over 10 % depending on the stage and type of tumor [41].

Most available methods use either whole genome or exome sequencing which require library preparation. Library preparation includes several sequential steps: 1) end-repair, 2) 5′-phosphorlyation; 3) addition of a terminal dA nucleotide to the 3′ ends of the fragments, 4) ligation of the fragments to adapters, 5) and polymerase chain reaction (PCR) amplification of the ligated products. Following PCR amplifications, the products are then quantified and sequenced. The reads are aligned and analyzed. Whole genome and exome sequencing have well documented biases resulting from GC content, DNA fragment size, quality, [42] and capture efficiency [43; 44].

This process can be simplified by sequencing long interspersed nuclear elements (LINEs). LINEs are non long terminal repeat retrotransposons that comprise 20% of the human genome and are present on every chromosome [45]. Because LINEs are highly repetitive, a discrete set of positions can be amplified from a single primer pair eliminating the need for end-repair, terminal 3′-dA addition, and ligation to adapters during library preparation. Reads can then be used to identify key types of structural and sequence variation important in human health. Using a predefined, discrete set of positions also has the potential to reduce alignment time, streamline the analysis

workflow, and achieve the necessary coverage at lower costs—all obstacles limiting the widespread use of whole genome and exome sequencing in genetic testing.

Numerous methods have been developed to identify structural changes from whole genome and exome sequencing reads (reviewed in [46]) but few have been developed specifically for lower admixtures rates typically observed from cell free DNA (reviewed in [47]). Most cell free DNA methods were designed for prenatal testing with much less work on cell free tumor DNA (reviewed in [48] [48]). Cancer screening can be much more challenging than prenatal testing due to the larger numbers of gains and losses observed across many different chromosomes [48].

My thesis work develops novel computation approaches to interpret insertion/deletion variants and detect aneuploidy and MSI using LINE sequencing.

Chapter 2: Variant Effect Scoring Tool for Insertions and Deletions (VEST-indel)

Insertion/deletion variants (indels) alter protein sequence and length, yet are highly prevalent in healthy populations, presenting a challenge to bioinformatics classifiers. Commonly used features—DNA and protein sequence conservation, indel length, and occurrence in repeat regions—are useful for inference of protein damage. However, these features can cause false positives when predicting the impact of indels on disease. Existing methods for indel classification suffer from low specificities, severely limiting clinical utility. Here, I further develop the Variant Effect Scoring Tool (VEST) to include the classification of in-frame and frameshift indels (VEST-indel) as pathogenic or benign.

2.1 History of Computational Assessment of Insertion and Deletion Variants

Most computational methods for assessing genetic variation initially focused on missense variants; more recently, several groups have extended these methods to handle indels [49-53]. Most of these methods utilize supervised machine learning classifiers and are trained on two classes of indel: pathogenic from disease mutation databases and benign from either population variation databases or tolerated interspecies variations derived from genomic alignments. DDIG-in is based on a support vector machine, and the authors of this method reported a sensitivity of 0.86 and specificity of 0.72 for frameshift indels [53], and a sensitivity of 0.89 for in-frame indels [50]; the authors did not report

prediction specificity for in-frame indels. PROVEAN uses an unsupervised approach that compares the reference protein sequence with a sequence that incorporates a variant of interest [49]. The authors of PROVEAN reported high sensitivities of 0.93 and 0.96 for in-frame insertions and deletions, respectively, and a specificity of 0.80 for in-frame insertions and 0.68 for in-frame deletions; PROVEAN does not assess frameshift indels. SIFT-indel, based on a J48 Decision Tree [54] , achieved good balanced accuracies for in-frame (sensitivity=0.81; specificity=0.82) [52] and frameshift indels (sensitivity=0.90; specificity=0.78) [51]. However, the neutral dataset used in those studies comprised indels derived from cross-species comparisons. As the authors state, SIFT-indel was trained to predict impact on gene function, irrespective of impact on disease. Indeed, when the method was applied to variants from human variation databases, the majority of the indels were predicted to be deleterious; thus, specificities would be below 50% for predicting indel pathogenicity. The CADD classifier utilized a unique approach, in which a support vector machine was trained to discriminate fixed (or nearly fixed) derived alleles in humans from a set of simulated variants [55]. The CADD classifier was developed to predict deleterious variants rather than variant pathogenicity or impact on protein function, but with the stated assumption that these quantities are all related. The authors of CADD reported classifier performance on missense variants and indels together, but not on indels separately [55].

2.2 VEST-indel

2.2.1 Data Collection

A curated set of in-frame and frameshift indels (micro-deletions and micro-insertions) of ≤20 base-pairs in length, annotated as being pathogenic from publications in the biomedical literature, was downloaded from Human Gene Mutation Database [56] (2014v.3). Only high-confidence annotations with the "DM" designation were included. A second curated set of in-frame and frameshift indels was downloaded from the NCBI ClinVar database on August 7, 2014. Only entries annotated as 'likely pathogenic' (Clinical Significance 4) or 'pathogenic' (Clinical Significance 5) and not annotated as a somatic mutation were included. Any entry from ClinVar that was also present in HGMD was removed from the ClinVar set. Annotated in-frame and frameshift variants were downloaded from the Exome Variant Server using (ESP6500SI-V2-SSA137) [57] [Fu, et al., 2013] and from the 1000 Genomes Project Phase 3 (ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/) [58]. To increase the likelihood that variants from the Exome Variant Server and 1000 Genomes Project were benign common polymorphisms, and to retain sufficient variants for the training set, I only used variants with a minor allele frequency (MAF) ≥ 0.01 and occurring in either African individuals or those of African ancestry. In ESP600, these were identified as "African-American" and in 1000G as the AFR superpopulation comprising YRI (Yoruba in Ibadan, Nigeria), LWK (luhya in Webuye, Kenya), GWD (Gambian in Western Divisions in the Gambia), MSL (Mende in Sierra Leone), ESN (Esan in Nigeria), ASW (Americans of African ancestry in SW USA), and ACG (African Caribbeans in Barbados). The other populations represented in ESP6500 and 1000G are believed to have experienced severe bottlenecks in recent

history, and hence individuals from these populations may harbor potentially pathogenic variants at higher MAF than individuals of African ancestry [51; 59-61]. A curated set of putatively benign in-frame and frameshift indels, derived from pairwise genome alignments of human and cow, dog, horse, chimpanzee, rhesus macaque and rat, was generously provided to us by Pauline Ng and Jing Hu. This set had been previously used to train their SIFT-indel classifier [51; 52]. Additional background information about these data sets, including probability densities for indel length and MAF, are shown in **Figure S 1**.

The number of variants used for this study, grouped by source and ontology, were: 2,523 in-frame deletions, 565 in-frame insertions, 17,606 frameshift deletions, and 8,265 frameshift insertions from HGMD 2014.3; 43 in-frame deletions, 14 in-frame insertions, 344 frameshift deletions, and 134 frameshift insertions from HGMD 2014.4; 1,991 in-frame deletions, 404 in-frame insertions, 774 frameshift deletions, and 618 frameshift insertions from ESP6500; 86 in-frame deletions, 70 in-frame insertions, 37 frameshift deletions, and 23 frameshift insertions from 1000 Genomes, Phase 1; 304 in-frame deletions, 261 in-frame insertions, 229 frameshift deletions, and 134 frameshift insertions from 1000 Genomes, Phase 3; 16 in-frame deletions, 5 in-frame insertions, 32 frameshift deletions, and 74 frameshift insertions from ClinVar; 4,686 in-frame deletions, 3,406 in-frame insertions, 706 frameshift deletions, and 628 frameshift insertions from the above-mentioned genome alignments.

2.2.2 Feature Selection

The Random Forest Feature Importance Z-score [62] was used to rank a set of 49 candidate features from [63] and 5 additional features (**Table S 1**), using PARF software (http://code.google.com/p/parf), with 100 trees and default parameters. To avoid overfitting, an independent feature-selection set was used (500 pathogenic and 500 benign examples for each of the in-frame and the frameshift classifiers). I used a greedy algorithm to identify a good, minimum set of features. Briefly, beginning with the top-ranked feature, a Random Forest was trained using only that feature and 10-fold cross-validation was used to estimate the classifier's area under the ROC curve (AUC). I successively added the next top-ranked feature until all candidate features were included. For the in-frame classifier, the maximum AUC was achieved with 23 features and for the out-of-frame classifier, the maximum AUC was achieved with 16 features (Table S 2). These features were used for the remainder of the work described here. The selected features include measures of gene importance, the damaging effect of the variant on protein activity, evolutionary conservation and protein local environment (Table S 3).

2.2.4 Classifier Training Protocol

Random Forest classifiers were trained to classify in-frame and frameshift variants (using PARF software with 100 trees and default parameters). For in-frame classifier training, 2,475 pathogenic and 1,877 benign examples were available, whilst 24,478 pathogenic and 1,350 benign examples were available for frameshift classifier training (**Table 1**). To handle class imbalance, the in-frame classifier was trained on a randomly selected set of 1,877 pathogenic examples and all 1,877 benign examples. Ten frameshift classifiers were trained on a randomly selected set of 1,350 pathogenic

examples and all 1,350 benign examples (repeated 10 times, sampling without replacement). All ten classifiers were used to score frameshift variants, by computing ten scores for each variant and averaging them.

Table 1: Datasets Used in Development of the VEST-Indel Method

	Feature selection	Training	Testing	Null distribution	Multi-Method Benchmark
In-frame					
Pathogenic	500[a]	2475[a]	39[b]	N/A	57[f]
Benign	500[c]	1877[c]	8105[d]	346[e]	156[e]
Frameshift					
Pathogenic	500[a]	24478[a]	184[b]	N/A	478[f]
Benign	500[c]	1350[c]	1340[d]	537[e]	60[e]

Superscript letters indicate the source of the examples for each type of insertion/deletion variant and each stage of VEST-indel development (feature selection, classifier training, classifier validation, empirical null) [a]HGMD, [b]ClinVar, [c]ESP6500, [d]Inter-species benigns from SIFT-indel, [e]1000G Phase 3, and [f]HGMD2014. There is no overlap between examples in any of the columns. N/A = not applicable because only benign examples were used to develop the empirical null distribution.

2.2.5 Statistical Framework

I developed an analytical null score distribution based on Random Forest classifier scores of putative benign variants (1000 Genomes Project Phase 3, MAF ≥0.01, African ancestry). The scored variants (537 for in-frame insertion/deletions, 346 for frameshift insertion/deletions) did not overlap the examples used for Random Forest feature selection, training or the independent test set. An empirical cumulative distribution (ECDF) of scores was calculated and modeled as a Generalized Pareto Distribution (GPD) (**Equation 1**)[64; 65].

Equation 1

$$F(z) = \begin{cases} 1 - \left(\frac{kz}{a}\right)^{\frac{1}{k}}, k! = 0 \\ 1 - e^{-\frac{z}{a}}, k = 0 \end{cases}$$

where k and a are the GPD shape and scale parameters, respectively (in-frame k=0, a=-8.48, frameshift k=0, a=-9.04) and z is the score ($0 \leq z \leq 1$).

Chapter 3: VEST-indel Performance Evaluation

In protein-coding exons, in-frame indels generally have a less severe impact than frameshifts [61]. Since the biological effect of in-frame and frameshift indels is different, I chose to develop two distinct Random Forest classifiers to handle these two distinct variant types. Performance of the classifiers was assessed in three phases: 1) I estimated sensitivity and specificity with stringent, homology-restricted ten-fold cross-validation (pathogenic class from HGMDv2014.3 [56], benign from ESP6500 African Ancestry [57]) (**Table 2**); 2) I re-estimated sensitivity and specificity on an independent test set of variants (pathogenic class from ClinVar [66], benign Interspecies alignments [51; 52]) (**Table 2**) that had not been used in classifier training and had been filtered for homology overlap with the cross-validation set; 3) I re-estimated sensitivity and specificity on a second independent test set of variants (pathogenic class from new entries in HGMDv2014.4, benign from 1000 Genomes Phase III African Ancestry [24] that did not overlap with any training data used by previously published methods (*multi-method benchmark set*); 4) I tested all possible meta-predictors that can be obtained from combining the four different indel classifiers using Boolean conjunctions and disjunctions. Finally, I sought to use VEST-indel scores to perform combine prioritization of different types of genetic variation (missense, in-frame, and frameshift variants).

Table 2: Training and Validation Sets Used by Current Prediction Methods

	Training Set (as published)		Test Set (as published)		Non-overlapping Multi-Method Benchmark Set	
	Pathogenic	**Benign**	**Pathogenic**	**Benign**	**Pathogenic**	**Benign**
In-frame						
PROVEAN	Uniprot	Uniprot	HGMD 2011	1000G P1	HGMD2014.4	1000G P3 AA
DDIG-in	HGMD 2012	1000G P1	Uniprot	Uniprot	HGMD2014.4	1000G P3 AA
SIFT-indel	HGMD 2010	Interspecies	Uniprot	Uniprot	HGMD2014.4	1000G P3 AA
CADD	Simulated	Fixed Polymorphisms	ClinVar	ESP6500	HGMD2014.4	1000G P3 AA
VEST-indel	HGMD 2014.3	ESP6500 AA	ClinVar	Interspecies	HGMD2014.4	1000G P3 AA
Frameshift						
PROVEAN	N/A	N/A	N/A	N/A	N/A	N/A
DDIG-in	HGMD 2012	1000G P1	HGMD 2012	Interspecies	HGMD2014.4	1000G P3 AA
SIFT-indel	HGMD 2010	Interspecies	N/A	N/A	HGMD2014.4	1000G P3 AA
CADD	Simulated	Fixed Polymorphisms	ClinVar	ESP6500	HGMD2014.4	1000G P3 AA
VEST-indel	HGMD 2014.3	ESP6500 AA	ClinVar	Interspecies	HGMD2014.4	1000G P3 AA

1000G P1 *and* 1000G P3 *are variants from 1000 Genomes Phase 1 and 3, respectively.* Interspecies *benign variants derived from pairwise genome alignments of human and cow, dog, horse, chimp, rhesus macaque, and rat.* Uniprot *variants were obtained from the UniProtKB/Swiss-Prot "Human Polymorphisms and Disease Mutations" dataset (Release 2011_09), annotated as deleterious, neutral, or unknown based on keywords from the provided Uniprot descriptions.* AA = *African or African American Ancestry and* N/A = *not applicable.*

3.1 Homology Restricted Cross Validation

The in-frame indel Random Forest and each of the ten frameshift Random Forests were assessed for sensitivity, specificity and balanced accuracy, using a rigorous ten-fold cross-validation protocol. The same protocol was applied to the missense Random Forest to assess combined prioritization of all three mutation types. To avoid overestimating performance, I ensured that any examples from genes whose protein products had ≥ 35% sequence identity were included in the same fold. BlastP with default parameters [67] was used for pairwise alignment of protein sequences and sequence identity calculations [67]. Evidence suggests that homology-restricted cross-validation is important to avoid overly-optimistic estimates of pathogenicity classifier performance [68]. In the cross-validation experiments, VEST-indel achieved a sensitivity and specificity of 0.90 for in-frame indels (**Table 3**). Cross-validation performance for frameshift indels was slightly lower, with a sensitivity of 0.83, specificity of 0.88, and balanced accuracy of 0.85.

Table 3: VEST-indel Performance Metrics.

	Sensitivity	Specificity	Balanced accuracy
In-frame Cross Validation	0.90	0.90	0.90
In-frame Testing	0.80	0.85	0.82
Frameshift Cross Validation	0.83	0.88	0.85
Frameshift Testing	0.89	0.86	0.87

Training utilized 10-fold cross validation and pathogenic variants from Human Gene Mutation Database 2014.3 and benign examples from Exome Sequencing Project (minor allele frequency in African Ancestry ≥ 0.01). The test set consisted of pathogenic examples from ClinVar and benign examples derived from pairwise genome alignments of human and cow, dog, horse, chimp, rhesus macaque, and rat.

3.2 Independent Test Set

An independent set of examples, having no overlap with feature selection, training or empirical null sets, was constructed. I removed any test set examples whose protein products had ≥ 35% sequence identity with any training examples [67]. Pathogenic in-frame and frameshift mutations were taken from ClinVar [66] and benign in-frame and frameshift variants were taken from the interspecies set. All data were 'cleaned' so as to ensure that there was no overlap between these examples and the other three data sets.

3.3 Comparison of insertion/deletion variant pathogenicity predictors

3.3.1 Multi-Method Benchmark Set

Four previously published methods were selected for comparison with VEST-indel. Three of the methods (SIFT-indel, DDIG-in, CADD) [49-53] handle both in-frame and frameshift insertion/deletions and the fourth method PROVEAN [49] handles only in-frame variants. To perform an unbiased comparison of VEST-indel and the four other methods, I identified a set of 553 pathogenic and 357 benign examples, which did not overlap with any examples used to train, fit parameters, select features, or validate performance by any of the four methods. This *multi-method benchmark set* comprised pathogenic examples (61 in-frame and 491 frameshift) from the most recent version of HGMD (2014v.4), excluding any examples present in earlier versions of HGMD that had been used to train DDIG-in, SIFT-indel, or VEST-indel (PROVEAN and CADD were not trained on HGMD). Benign examples (224 in-frame and 118 frameshift) were taken from 1000G Phase 3 (MAF ≥0.10, African Ancestry). Any examples present in 1000G

Phase 1 or ESP6500, which were used to train or validate any of PROVEAN, DDIG-in, CADD or VEST-indel were omitted. Only examples for which every method returned a prediction result were included. The final multi-method benchmark set comprised 59 benign frameshift insertion/deletion and 163 benign in-frame insertion/deletion variants (MAF≥0.1 from 1000G AFR super-population), as well as 474 pathogenic frameshift and 53 pathogenic in-frame variants from HGMD v.2014.4 (http://karchinlab.org/vest_indel_additional_files/Additional_File_2.xlsx).

3.3.2 Method Usage

SIFT-indel and DDIG-in provide a categorical classification for each example (Damaging/Neutral or Disease/Neutral) and a confidence measure. PROVEAN, CADD and VEST-indel provide a numerical score for each example (-40 to 12.5, 1 to 99, 0 to 1). To compare methods, PROVEAN scores were assigned to categories of Damaging (<-2.5) or Neutral (≥-2.5) (as recommended by the authors) and VEST-indel scores were assigned to categories of Pathogenic (≥0.5) or Benign (<0.5), which represents a majority vote of decision trees in the Random Forest classifier. CADD scaled "C-scores" were assigned to categories of Deleterious (<15) or not Deleterious (≥15) (as recommended on their webserver). Sensitivity (TP/(TP+FN)), specificity (TN/(TN+FP)) and balanced accuracy ((sensitivity+specificity)/2) were calculated for each method, where TP = the number correctly classified as pathogenic (or damaging or disease) examples, FN = the number of incorrectly classified pathogenic examples, TN = the number of correctly classified benign (or neutral) examples, and FP = the number of incorrectly classified benign examples.

3.3.3 Indel Classifier Performance

Table 4 compares performance achieved by the five methods for classifying neutral and disease-causing indels from the multi-method benchmark set. VEST-indel shows superior specificity for classifying both in-frame (0.96) and frameshift (0.95) indels. These high specificities further validate the ability of VEST-indel to accurately reject neutral variants as disease causing. All methods had reasonably high balanced accuracies for in-frame indel classification, with VEST-indel and PROVEAN yielding the highest balanced accuracy of 0.88. Of note, VEST-indel and PROVEAN achieved nearly identical balanced accuracies with approximately equal trade-offs in sensitivity and specificity (**Table 3**). For frameshift variants, VEST-indel outperformed the other methods, having a balanced accuracy of 0.90, compared with 0.77 for DDIG-in, 0.59 for SIFT-indel, and 0.52 for CADD. In the case of DDIG-in, VEST-indel showed substantially improved sensitivity and specificity (Table 4). The dramatic gain in performance achieved by VEST-indel, relative to SIFT-indel and CADD, resulted from a marked gain in specificity (0.95 vs. 0.25 for SIFT, and 0.05 for CADD); this is consistent with previous reports for SIFT-indel, which maintains good specificity when predicting protein-damaging indels, but suffers low specificity when predicting pathogenicity }[51; 52].

Table 4: Comparing performance with previously published results and testing all methods with the new multi-method benchmark dataset.

	Previously published		Multi-method benchmark		
	Sensitivity	**Specificity**	**Sensitivity**	**Specificity**	**Balanced Accuracy**
In-frame					
VEST-indel	0.90†	0.90†	0.81	0.96	0.88
SIFT-indel	0.81	0.82	0.86	0.76	0.81
DDIG-in	0.89	N/A	0.78	0.91	0.84
PROVEAN	0.93/0.96	0.80/0.68	0.95	0.80	0.88
CADD	N/A	N/A	0.74	0.88	0.81
Frameshift					
VEST-indel	0.83†	0.88†	0.85	0.95	0.90
SIFT-indel	0.90	0.78	0.94	0.25	0.59
DDIG-in	0.86	0.72	0.75	0.80	0.77
CADD	N/A	N/A	0.98	0.05	0.52

Previously published sensitivity and specificity based on author's cross-validation experiments. PROVEAN does not use cross validation so the reported numbers are from validation set experiments done separately for insertion and deletion variants. *N/A* = not applicable. Published results for the DDIG-in in-frame classifier do not include specificity; their self-reporting consists of an accuracy (not balanced accuracy) of 0.84 and precision of 0.81. The authors of CADD did not report the performance achieved with indels separately. †Results from **Table 1**included here for comparison. Multi-method benchmark set consisted of pathogenic examples from Human Gene Mutation Database 2014.4 and benign examples 1000 Genomes Phase 3 (minor allele frequency in African Ancestry ≥ 0.1).

3.4 Boolean Meta-predictors

In these Boolean expressions, each method is represented by a variable X_i, which is set to TRUE when the method classifies an example as pathogenic and FALSE when the method classifies an example as benign. For combinations of two methods, candidate meta-predictors were $(X_1 \text{ AND } X_2)$ and $(X_1 \text{ OR } X_2)$. For combinations of three methods, candidate meta-predictors $(X_1 \text{ AND } X_2 \text{ AND } X_3)$, $(X_1 \text{ OR } X_2 \text{ OR } X_3)$, $(X_1 \text{ OR } X_2 \text{ OR } X_3)$, $((X_1 \text{ AND } X_2) \text{ OR } X_3)$, $((X_1 \text{ OR } X_2) \text{ AND } X_3)$, $((X_1 \text{ AND } X_3) \text{ OR } X_2)$, $((X_1 \text{ OR } X_3) \text{ AND } X_2)$, $((X_2 \text{ AND } X_3) \text{ OR } X_1)$, $((X_2 \text{ OR } X_3) \text{ AND } X_1)$. For combinations of four methods, there are 64 possible combinations (**Table S 4**). I used a brute-force approach and limited the number of methods in the meta-predictor to a maximum of four to avoid a combinatorial explosion. All possible four-way combinations of the five methods were explored.

Although VEST-indel, SIFT-indel, DDIG-in, PROVEAN and CADD share some similarities with respect to training sets and features, I considered that they might be different enough to provide independent information about an indel of interest. Therefore, they could be combined into a meta-predictor to yield improved performance. This approach has had some success in predicting the pathogenicity of missense variants [69-71]. Using the multi-method benchmark set, I assessed the classification performance resulting from each pair, trio, or quartet of methods combined using Boolean conjunctions and disjunctions. See Supp. Table S6-7 for a complete list of the tested combinations.

For in-frame classification, the combination of ((VEST-indel AND PROVEAN) OR (CADD AND DDIG-in)) yielded a substantially improved sensitivity (0.93) while retaining good specificity (0.97), when compared to VEST-indel alone (sensitivity=0.81, specificity=0.96), and indeed any of the methods alone (**Table S 6**). This result indicates that these methods are highly complementary when combined in the described fashion. Conversely, for frameshift classification, the combination of ((VEST-indel AND (SIFT-indel OR DDIG-in)) had roughly equivalent sensitivity (0.83) and specificity (0.97) to VEST-indel alone (sensitivity=0.85, specificity=0.95). This results because the most specific method (VEST-indel) is combined using the AND operation (i.e., sensitivity could not possibly increase, nor could specificity decrease).

The strategy of classifying a variant as pathogenic if any of the classifiers predicted it to be pathogenic (i.e., combining classifiers with a Boolean OR) did not yield good results. For the in-frame classifier, the combination (VEST-indel OR SIFT-indel OR PROVEAN OR CADD) had a sensitivity of 1 but a specificity of 0.56, with balanced accuracy of 0.78. Combining four classifiers or three classifiers with the OR operator consistently yielded good sensitivity but a substantial decrease in specificity. This result is, to some extent, expected because combining classifiers with the OR operation increases the possibility of accepting a variant as pathogenic. Conversely, requiring that all classifiers agree (i.e., combining classifiers with a Boolean AND) reduces the probability of a pathogenic classification. Indeed, all meta-predictors that used only AND operators had high specificity, but low sensitivity. For example, the (VEST-indel AND SIFT-indel AND CADD AND DDIG-in) meta-predictor had a specificity of 1.00 and sensitivity of 0.46. Taken together, these results highlight the benefit of developing

meta-predictors that combine Boolean conjunctions and disjunctions, rather than considering only a single type of Boolean operation.

3.5 Combined Prioritization of Indel and Missense Variants

For each in-frame, frameshift or missense variant, a VEST score was computed using homology-restricted 10-fold cross-validation with the appropriate Random Forest (Performance Assessment). Then a p-value was calculated using the analytical null (**Equation 1**) for its respective type. To assess whether VEST could correctly prioritize a pathogenic variant over a benign variant, irrespective of whether the variant was in-frame, frameshift or missense, I ranked the combined set of variants according to p-value, and computed area under the Receiving Operator Characteristic (ROC) curve [72].

To compare VEST results with CADD, which also provides combined prioritization, I scored the same variants using the CADD. Variants were ranked according to their scaled C-scores and area under the ROC curve was computed.

VEST-indel p-values for in-frame and frameshift indels are comparable to VEST p-values for missense variants and as a result, multiple variant types can be jointly prioritized. I assessed joint prioritization performance by combining variants from the VEST-indel in-frame and frameshift training sets (**Table 1**) and variants from the VEST missense training set [73] (2475 pathogenic and 1877 benign in-frame indels; 24478 pathogenic and 1350 benign frameshift indels; 38221 pathogenic and 38221 benign missense variants) (**Figure 1**). I also assessed performance in a balanced set, in which I randomly selected 1350 pathogenic and 1350 benign variants of each type for the combined set. VEST p-values and scaled CADD scores were used to compute ROC area under the curve (AUC). For the combined set, VEST and CADD achieved a similar

ROC area under the curve (AUC) of 0.90 and 0.88, respectively. For the balanced set, VEST classification resulted in an AUC of 0.91 and CADD classification resulted in an AUC of 0.74.

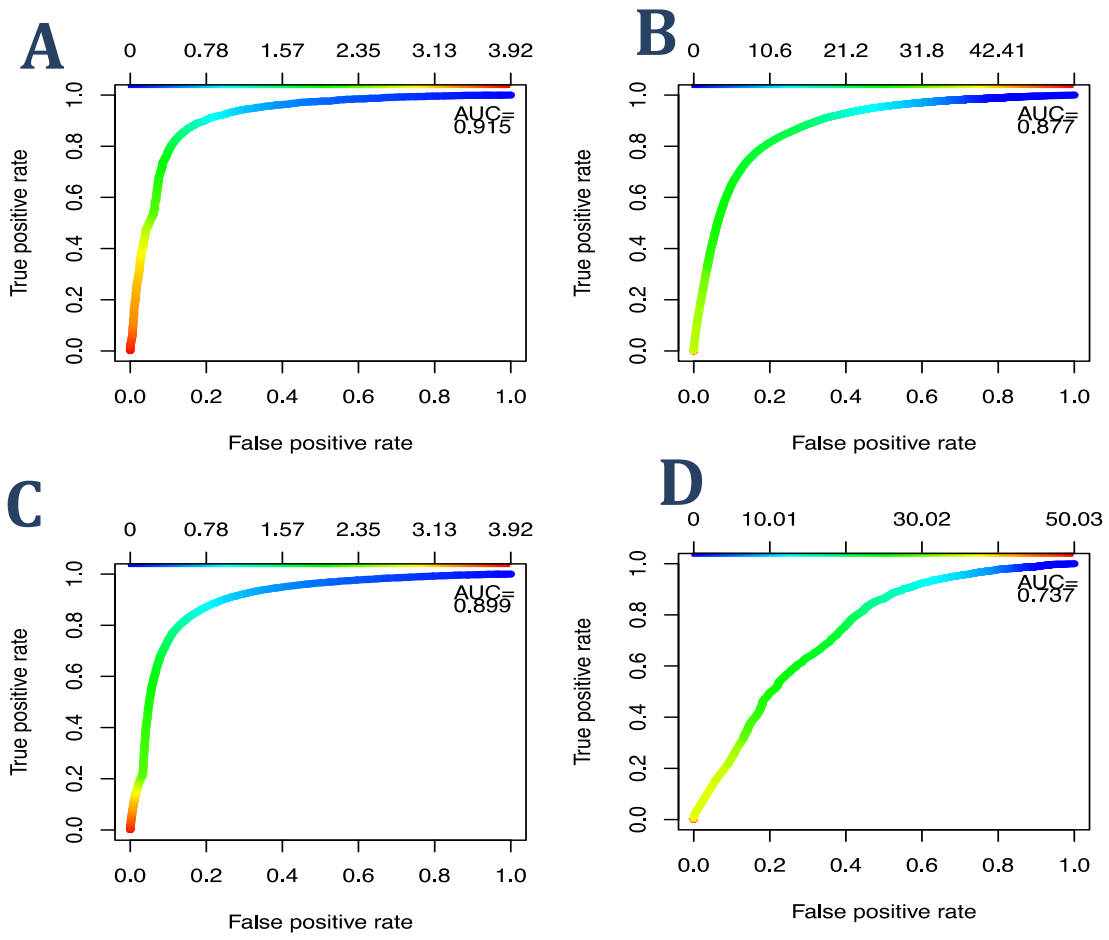

Figure 1: Combined Prioritization results comparing VEST and CADD.
A) VEST 3.0 ROC Curve on the variant test set dominated by missense variants B) CADD ROC Curve on the test set dominated by missense variants C) VEST 3.0 ROC Curve on the variant test set where by missense and indels have equal proportions D) CADD ROC Curve on the test set where by missense and indels have equal proportions

3.6 Discussion

3.6.1 Selective pressures on genes and false positive classifications

A standard method for identifying genes under selection is Tajima's D statistic[74], and for each gene harboring a variant in the multi-method benchmark set, I computed this statistic based on its longest annotated RefSeq transcript [75]. If RefSeq transcripts were not available for the gene, the longest annotated Ensembl transcript [76] [Cunningham, et al., 2015] was used. These calculations were performed using SNPs in 1000 Genomes Phase 3 AFR samples and the PopGenome package in R [77] [Pfeifer, et al., 2014]. Each gene was assessed for the presence of statistically significant positive or balancing selection ($p<0.05$). The PopGenome package estimates P-values by simulation using Hudson's coalescent model [Hudson, 2002]. The benign examples in the multi-method benchmark set were taken from the 1000 Genomes Phase 3 samples, limited to individuals in the AFR (African) super-population [24] and having MAF≥0.1. Whereas common variants are generally considered to be non-pathogenic [78], datasets of common variants may be contaminated by pathogenic variants if they occur in genes that are not under purifying selection [51]. I assessed the possibility that the multi-method benchmark set might include common pathogenic variants. If this were the case, a false positive call from one of the methods might represent the correct identification of a truly pathogenic variant. Genes not subject to purifying selection might alternatively be under positive, balancing, or relaxed (neutral) selection [51]. For each of VEST-indel, SIFT-indel, DDIG-in, PROVEAN, and CADD I assessed the relationship between variants under selective pressure and those called as false positives, using Fisher's exact test (two-tailed, $\alpha=0.05$). None of the benign variants were under balancing selection, defined as a statistically

significant (nominal $p<0.05$) positive Tajima's D statistic. Thirteen frameshift and 56 in-frame variants were under positive selection, defined as a statistically significant (nominal $p<0.05$) negative Tajima's D statistic. With the exception of a borderline p-value for DDIG-in in-frame variant classification ($p=0.051$), there were no statistically significant relationships between positively selected variants and variants that were called as false positives. For frameshift variants, $p=1.0$ for VEST-indel, $p=0.26$ for SIFT-indel, $p=0.18$ for DDIG-in, and $p=0.40$ for CADD. For in-frame variants, $p=0.25$ for VEST-indel, $p=0.56$ for SIFT-indel, $p=0.13$ for PROVEAN, and $p=0.79$ for CADD.

3.6.2 Conclusion

In this study, I sought to develop a method for predicting indel pathogenicity. This functionality is distinct from existing classifiers that were developed to predict indel impact on protein structure or function. Although clinical utility appears to be a common goal for much of bioinformatics methods development, indel pathogenicity prediction presents the challenge of distinguishing variants that affect protein structure and function from those that adversely affect health [24]. Given the enrichment for protein sequence and annotation features available for algorithmic development [79; 80], the difficulty discriminating neutral and disease-causing indels might be unsurprising. The newly developed classifier, VEST-indel, partially addresses previous methodological limitations, and achieves high balanced accuracy even when tasked with sorting disease-associated indels from those present in the general population. In particular, VEST-indel realized substantial gains in specificity relative to existing methods, highlighting reductions in falsely classifying neutral variants as pathogenic. To realize these performance gains, VEST-indel heavily utilized a new feature that captures the known

relevance of a gene to human health. This new "PubMed" feature leverages decades of community-wide biomedical research. Thus, the algorithm uses features that ultimately estimate indel impact on protein, and the PubMed feature additionally estimates the biological context of the protein. Given that poor specificity also limits the utility of methods aimed at assessing the pathogenicity of missense variants [81-84], the approach presented here might prove beneficial for variant classification in general.

The in-frame meta-predictor ((VEST-indel OR DDIG-in) AND (PROVEAN)) achieved excellent sensitivity (0.93) and specificity (0.94) when applied to the multi-method benchmark dataset. This complementarity results because the two high-specificity classifiers are combined using the OR operation, which is then combined with the high-sensitivity classifier PROVEAN, using the AND operation (see individual classifier performance, **Table 4**). The Boolean OR operation increases the possibility of pathogenic classification; importantly, pathogenic classification from VEST-indel and DDIG-in is complementary rather than entirely overlapping, hence the increased sensitivity relative to either method alone. As expected, however, the specificity of the (VEST-indel OR DDIG-in) classifier decreased (see **Table S 6**). Next, even though the highly sensitive PROVEAN is slightly more prone to false positives, the specificity of the meta-predictor cannot decrease owing to the unanimity required by the AND operation; on the contrary, the complementarity of true-negative calls among these three classifiers restores a high specificity. This is deliberate in the explanation because meta-predictor derivation relying on a single Boolean operation type is limiting and can result in significant trade-offs in sensitivity and specificity. As the results show, taking advantage

of the complementarity that can result from combining Boolean conjunctions and disjunctions can be beneficial when maximizing balanced accuracy is desired.

The new VEST-indel method can be used in combination with VEST scoring of missense variants to yield a jointly prioritized list of both variant types. This analysis requires a single batch submission to the CRAVAT server [85]. To my knowledge, the only other automated method available for such joint ranking is CADD. For data sets in which the number of missense variants far exceeds the number of indels, VEST and CADD have similar performance. However, when variant types (indels and missense) and classes (pathogenic vs. benign) are evenly distributed, VEST significantly outperforms CADD.

Chapter 4: Aneuploidy Detection

Over 100 years ago, Theodor Boveri demonstrated sea urchins require the proper number of chromosomes for embryonic development [30]. He also noted the role of abnormal number of chromosomes in cancer [30]. Developing techniques to accurately detect aneuploidy is important for both prenatal screening and cancer detection.

4.1 Prenatal Screening

Approximately 1 of 154 live births have a major chromosomal abnormality [86]. This rate increases with the inclusion of still-born or miscarriage pregnancies [87]. Some of the most common chromosomal abnormalities include Trisomy 21 (Down syndrome) 1 in 831 live births, Triple X 1 in 909, XYY 1 in 969, XXY 1 in 969, Trisomy 18 (Edward's Syndrome) 1 in 7,573, and Trisomy 13 (Patau syndrome) 1 in 22,719 [86]. Maternal age plays a significant role in chromosomal abnormalities. At age 35 the rate of chromosomal abnormalities in live births is 0.52% but increases to 5.57% by age 45 [86]. Despite the high rate of abnormalities, it is estimated that 30% of all chromosomal abnormalities will not be detected [88] hence the need for screening techniques.

Prenatal screening protocols such as chorionic villus sampling and amniotic fluid sampling are the current gold standard [89] but these invasive methods have a miscarriage rate of 1 out of 100 [33]. Noninvasive prenatal testing (NIPT) using cell free fetal DNA (cffDNA) in maternal plasma has many advantages over invasive techniques. NIPT no longer has a risk of fetal loss and greatly reduces pain and complications [34]. In a meta-analysis of 37 studies, NIPT outperforms traditional invasive methods for Trisomy 21 with a sensitivity of 99.2% and specificity of 99.91% while performance for other trisomies (13, 18, sex) was slightly lower [90]. Initial studies were largely conducted using

pools of high-risk pregnancies which have a higher probability of chromosomal abnormalities than the general population [47]. One study examined NIPT performance in the general population and found NIPT protocols had lower positive predictive power but still outperformed traditional invasive protocols [91]. Despite the many advantages of NIPT, there are several obstacles before NIPT supplants traditional invasive procedures including: cost, protocol accessibility, and cytogenetic diagnostic performance on other abnormalities [47].

4.2 Cancer Detection

Aneuploidy is present in most types of cancers [92] and malignancy is directly proportional to amount of abnormalities [32]. Unlike single chromosomal abnormalities observed in fetal screening, cancers typically have multiple chromosomal abnormalities [93]. Not all chromosomal arms are likely to be altered and gains/losses do not occur in the same frequency. The most frequently observed gains are: 1q, 5p, 7p, 8q, 20q while the most common losses are 4q, 6q, 8p, 13q, 17p [93]. There is a pressing need to detect chromosomal abnormalities in cancer.

Invasive biopsies are the gold standard for retrieving cancer tissue for cancer diagnosis. Biopsies, however, can cause severe pain [34] and occur in tissue types that are difficult to sample. Like NIPT, genetic screening using cell free DNA has many advantages over invasive techniques. Tumor DNA (cftDNA) can be retrieved from blood, urine, and stool and comprises 0.01% to 90% of all circulating DNA [35; 94]. Malignancies with as few as 50 million cells can be detected which is far below the resolution of radiological imagining and suggests that cftDNA could be used early cancer screening [94].

4.3 Current Techniques to Detect Aneuploidy

4.3.1 G-Banding

G-Banding was the first method developed to detect chromosomal abnormalities [95]. In this procedure, cells in mitosis are treated with colchicine causing the microtubules to disrupt and arrest in metaphase. Cells are then fixed and treated on a glass microscope slide with Giemsa dye. The dye produces distinct and reproducible patterns on each chromosome for study.

4.3.2 Fluorescence *In Situ* Hybridization (FISH)

Fluorescence *In Situ* Hybridization (FISH) attaches a fluorescently labeled DNA probe that can bind to chromosomal regions with sequences are highly complementary. Researchers use fluorescence microscopy to identify chromosomal locations where the bound probe resides [96]. This protocol can be expanded to include different color probes for use on all chromosomes and is often referred to as multiplex-FISH [97; 98].

4.3.3 Comparative Genome Hybridization (CGH)

Comparative Genome Hybridization isolates and fragments DNA for both a test and control sample. One sample is labeled with red fluorescent dye and the other sample is labeled with green dye. The dyed fragments compete and bind to different probes. Researchers can assess whether a region is amplified, deleted, or unaltered based on the whether the corresponding chromosomal region appears more green or red. The time intensive standard CGH protocol was extended to include thousands or base-pair fragment on a microchip and the resulting color analysis has been automated (microarray). The information from a single microarray experiment is more than thousands of FISH experiments [99].

4.3.4 Digital Karyotype

Digital Karyotype quantifies short sequences of DNA from specified loci over the genome. Aneuploidies can be identified based on the read depth at a loci of interest. Methods have been designed to identify both small and very large aneuploidies, with a wide range of techniques including circular binary segmentation, hidden Markov models, expectation maximization and mean-shift (as reviewed in [46]).

The initial focus of read depth methods was whole genome sequencing, in which reads are expected to be randomly distributed across the genome. Under the assumption that reads are uniformly and independently distributed, regions of normal copy number are expected to follow a Poisson or Normal distribution [46; 100]. These approaches require correction for biases induced by differing GC content across the genome and uneven representation of genomic regions in library preparation [46]. Read depth methods have subsequently been extended to targeted, capture-based sequencing protocols, requiring modification of segmentation algorithms and introduction of alternate techniques, such as regression models and PCA. Targeted sequencing introduces coverage discontinuities, and increased variability due to differences in capture efficiency [46]. Several recent methods also leverage randomly distributed off-target reads to improve accuracy [101; 102].

Amplicon-based sequencing approaches also have utility for aneuploidy detection, but fewer aneuploidy analytical methods have been developed for this purpose. Amplicon-based protocols can achieve high coverage depth at relatively low cost, and they are an attractive alternative to WGS and targeted protocols for small amounts of DNA at low neoplastic fractions [103; 104]. Reads from amplicon sequencing are limited to a relatively small number of discrete loci. In addition to being discontinuous, they are not

randomly distributed, which makes it difficult to apply existing mathematical models of read depth coverage. Several studies have reverted to Z-score approaches in which read depth at a locus is compared to an overall mean, and aneuploidy is called when the centered read depth exceeds 3-5 standard deviations away from the mean [105-107]. In a refinement of this approach, read depth may be converted to a ratio between test sample and a control at the locus, using either a matched normal or a pool of selected control samples [108]. In the next chapter, I present a new analytical approach for aneuploidy detection from FastSeqs amplicon sequencing of long interspersed nucleotide elements (LINEs). FastSeqs is fast and efficient but yields a read depth distribution that is not well handled by computational methods developed for WGS and targeted sequencing. I use a within-sample approach to identify chromosome arm level gains and losses and a machine learning method to summarize the general aneuploidy of a sample.

Chapter 5: WALDO: Within-sample AneupLoidy DiscOvery

Here, I introduce a novel computational framework entitled WALDO (Within-sample AneupLoidy DiscOvery) (**Figure 2**). This method contributes several methodological advances to amplicon-based aneuploidy detection. The method identifies groups of genomic intervals distributed throughout all chromosomes, whose read depths track together in DNA from normal individuals, because they have similar amplification properties. Although the reads are not randomly distributed across the genome, within these groups, they are approximately random and Normally distributed. For a test sample of interest, the method uses a within-sample approach to parameterize a Normal distribution (mean and variance) of read depths for each group. It is expected that the total read depth, summed over all genomic intervals in a euploid chromosome arm, will not deviate significantly from its theoretical expectation. A convenient property of the Normal distribution is that the theoretical mean and variance can be computed by summing the group means and the group variances of the chromosome arm intervals. To my knowledge, this is the first time that within-sample methods have been applied to amplicon sequencing.

I show that this framework can be applied to identify chromosome arm gains or losses, including allele-specific gains and losses. Furthermore, this method incorporates machine learning to make generalized aneuploidy calls, in which samples are classified according to their aneuploidy status. This study presents a significant advance over previous work. This protocol reduces the false positive rate with respect to calling

aneuploid chromosome arms, and I show that the method can be applied robustly on thousands of diverse samples, including primary biopsy samples of 11 different tumor types with high neoplastic fraction and samples with fractions as low as 1%. While very low neoplastic fractions have been previously reported, these studies have been limited to small sample size (<50 samples) [106; 109; 110] without the need to handle batch effect issues. Matched normal samples are not required for the aneuploidy calling method, but if matched normals are available for a test sample, this method can also call somatic single base substitution and indel mutations, estimate mutation load and identify carcinogen signatures and microsatellite instability (**Figure 2**).

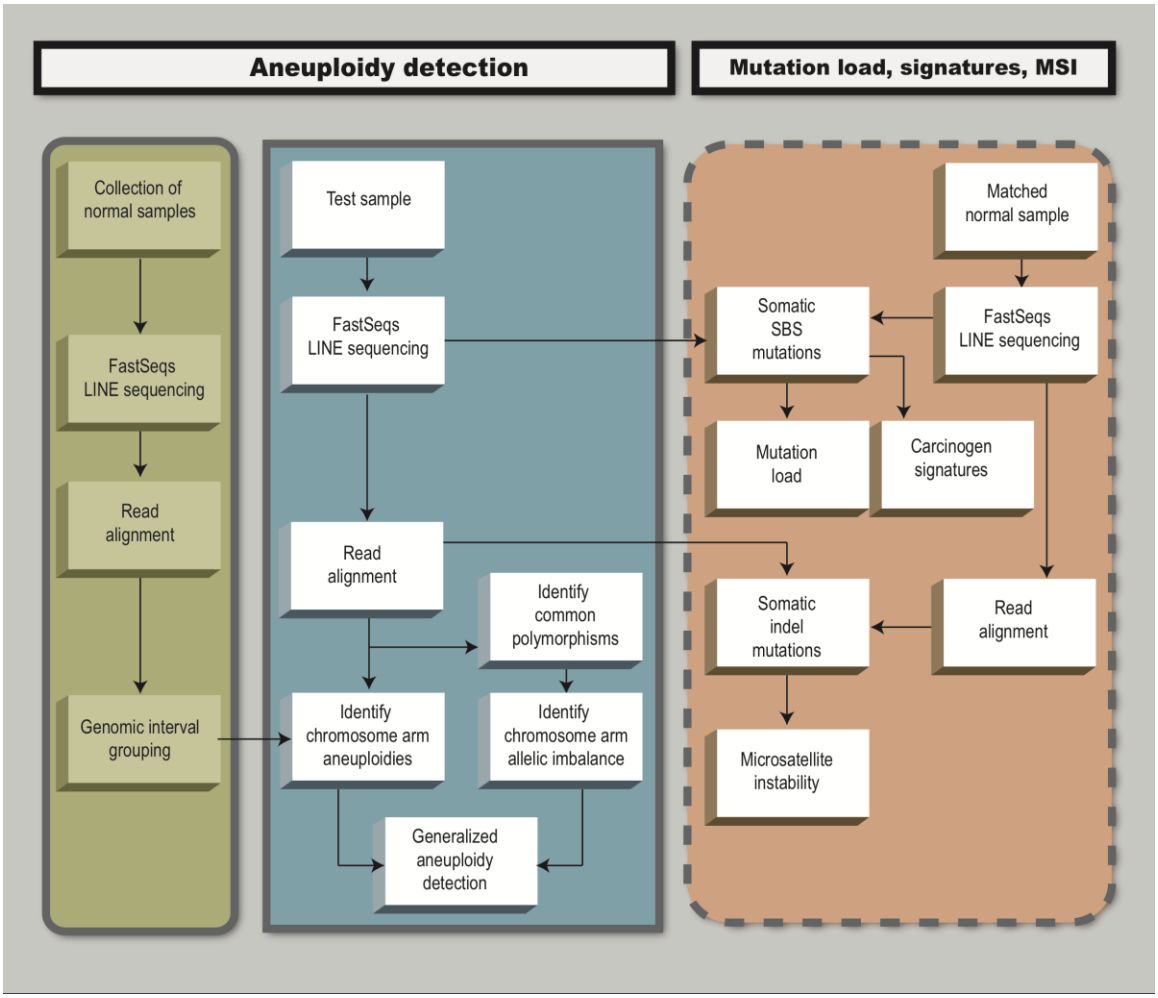

Figure 2: WALDO Overview
WALDO identifies chromosome arm gains or losses, including allele-specific gains and losses without the use of a normal. When matched normals are available for a test sample, this method can also call somatic single base substitution and indel mutations, estimate mutation load and identify carcinogen signatures and microsatellite instability

5.1 Sample Collection

Sample collection

562 primary tumors (16 of which had adjacent normal tissue), and 176 normal white blood cells (WBC) were collected. The primary tumors consisted of (45 bladder cancers, 56 breast cancers, 302 colon and colorectal cancers, 26 esophageal cancers, 37 head and neck, 27 liver cancers, 9 ovarian cancers, 22 uterine cancers, 22 gastric cancers, 10 urothelial carcinomas of the upper urinary tract, 6 colorectal cancers with microsatellite instability and 32 colorectal adenomas. Peripheral white-blood-cell (WBC) and plasma samples were collected from healthy individuals. All individuals provided written informed consent after approval by the institutional review board of The Johns Hopkins University.

5.2 Fast-SeqS and Safe-SeqS

For each sample, FAST-SeqS was used to amplify approximately 38,000 amplicons, with a single primer pair [105], and massively parallel sequencing was performed on Illumina instruments (HiSeq, MiSeq). During amplification, degenerate bases at the 5' end of the primer were used as barcodes to uniquely label each DNA template molecule (Safe-SeqS [111]. To reduce the possibility that a DNA template molecule would be counted multiple times, the barcodes were used to quantify unique templates (UIDs). The UIDs were used instead of raw reads. Samples were sequenced to a depth of 15-25 million reads using 3-15 million UIDs. Normal replicates were included in every sequencing run as a positive control, and replicates were used to capture differences resulting from stochastic and experimental variability.

5.3 Sample alignment and genomic interval grouping

For each sample, read length was 150bp and targeted read depth was 25M. Reads were grouped into families sharing a common SAFE-seqs barcode, yielding 3-15M unique reads (UIDs). Bowtie2 was used to align reads to human reference genome assembly GRC37. Next, 24,720 single nucleotide and 1,500 insertion and deletion polymorphisms with MAF>1% were extracted from 1000G [112] (Phase 3 20130502). I identified 37,669 exact matches (33,844 on autosomes) to the reference genome, allowing also for exact matches that included a common polymorphism. For efficiency. the genomic positions and amplicon sequences were stored and used to directly map amplicons from subsequent samples, without a sequence alignment step.

Even when there was no aneuploidy present, it is expected that the number of UIDs that mapped to a genomic region would be variable, based on stochastic and experimental factors. This variability was controlled for by grouping genomic intervals with similar UID depth across all chromosomes in *multiple normal samples*. Intervals consisted of 500Kb of genomic DNA, which spanned one or more LINEs, with a total 4361 regions across 39 non-acrocentric chromosomes. The 500Kb length produced good performance by the clustering algorithm. Increasing the length reduced its power to identify similar intervals and decreasing the length increased computational expense.

Genomic interval grouping was performed for each test sample by selecting a set of normal (non-aneuploid) samples which had similar distributions of DNA amplicon sizes to that sample. Briefly, during PCR, smaller amplicons are preferentially amplified [105; 113]. The distribution of amplicon sizes depends on both the batch in which a sample is

amplified and its source. For each test sample *p*, I selected the seven normal samples with the smallest Euclidean distance to *p (***Equation 2***),*

Equation 2

$$D(p,q)=\sqrt{\sum_{n}(q_n - p_n)^2}$$

where, p_n and q_n are the fraction of amplicons of size *n* in samples p and q, and the sum is over all amplicon sizes in the two samples. The variance of amplicon sequence depth across samples was estimated with maximum likelihood and the top 1% were removed. Additionally, any amplicons with < 10 UIDs in one sample and > 50 UIDs in any other sample were removed. Large differences among normal samples were most likely caused by private sequence variants that were not mapped to the correct amplicon. LINE-spanning genomic intervals were clustered by their UID depth distributions, after scaling, across the selected normal samples. Scaled UID counts were computed by subtracting the mean and dividing by the standard deviation of UID counts in each sample.

First, each chromosome interval ***i*** was assigned to a primary cluster $\boldsymbol{C_i}$. Next, UID depth distribution of ***i*** across all samples was compared to all other intervals ***i'*** that occurred on the remaining 21 autosomal chromosomes, across all samples. Since I was looking for similarity, I tested for insignificant results (paired t-test p>0.05, f-test p>0.05). If an ***i'*** was similar to ***i*** , it was added to cluster $\boldsymbol{C_i}$**.** This procedure allowed for an interval ***i*** to be the sole member of its primary cluster and/or belong to more than one cluster. The average cluster contained ~200 intervals (**Figure 3**). The basic protocol

uses normal samples only for genomic interval grouping, and all statistical tests are done based on UID count distributions in a test sample.

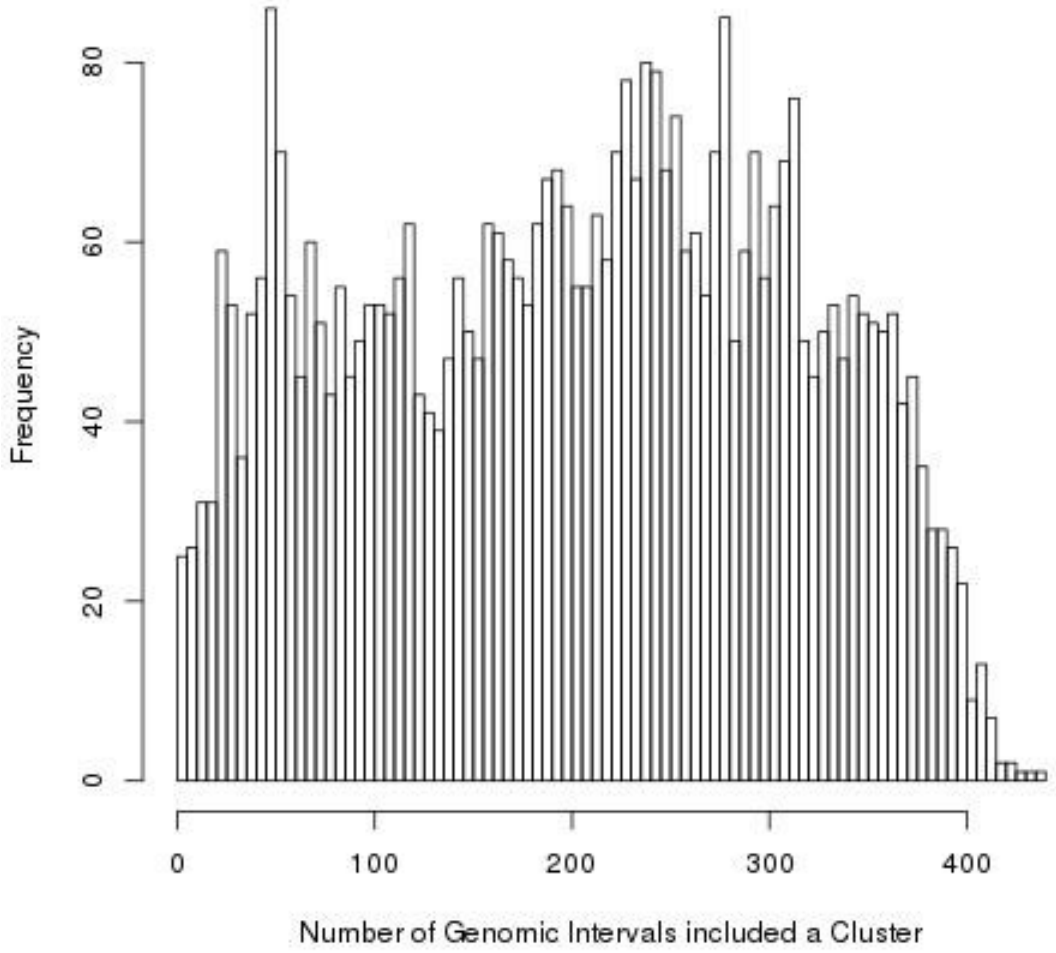

Figure 3: Representative Sample to illustrate the Number of Genomic Intervals included in a Cluster

5.4 Calling chromosome arm aneuploidies in a test sample

While FastSeqs UIDs were not randomly distributed, UIDs in each genomic interval group yielded by the protocol follow an approximately normal distribution (**Figure 4**). For a test sample, maximum likelihood estimation determines the UID depth mean μ and variance σ^2 of each the 4,361 genomic interval groups. This improved the robustness of these estimates by iteratively removing outliers from the groups. For each group, I flagged any outlier interval with (min(2* CDF(μ, σ_i^2), 2*(1-CDF)) 1-CDF(μ, σ_i^2) < α (α=0.01 in this work), and the flagged interval was removed from all groups. Next, the and parameters of each group were re-estimated by maximum likelihood. The two steps were repeated until no outlier intervals remained.

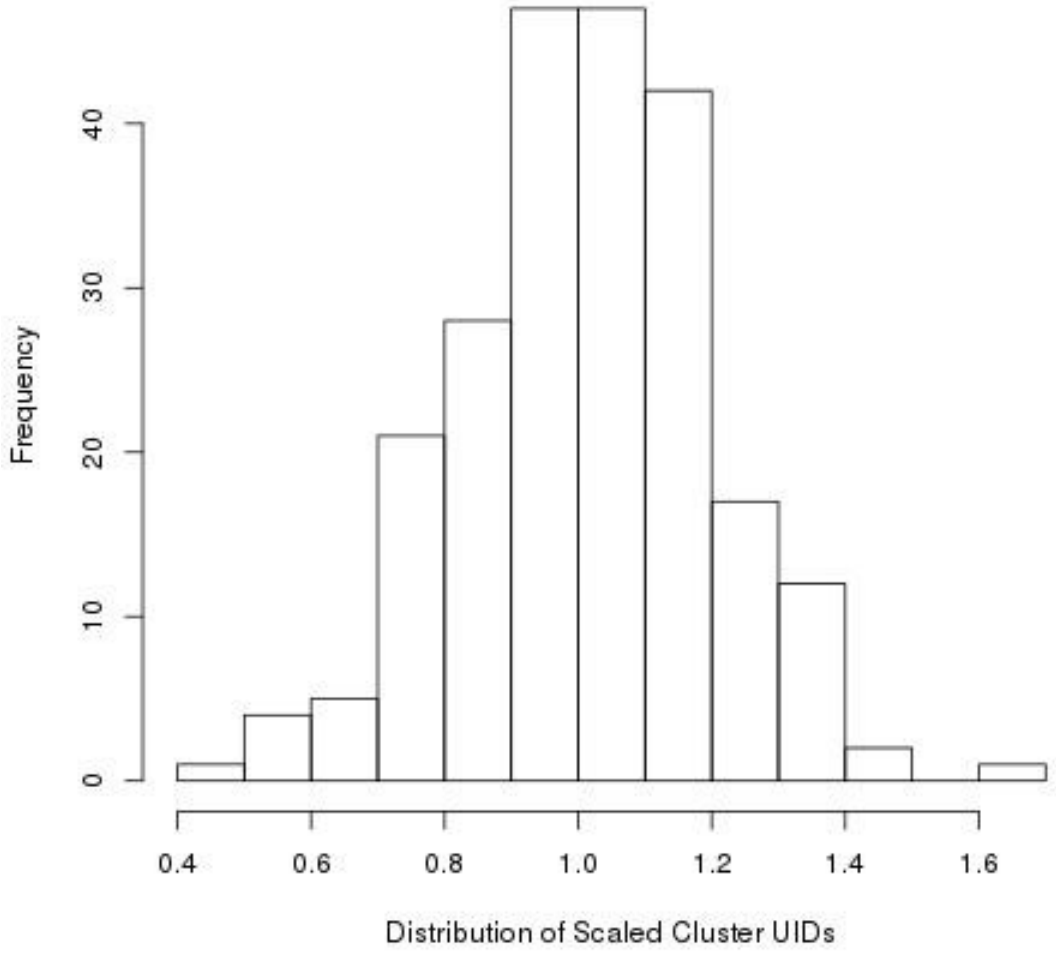

Figure 4: Distribution of Scaled Cluster UIDs in a Representative Cluster

To infer whether a chromosome arm was aneuploid, I estimated the statistical significance of the total UID count of the arm, by comparing to the μ and σ^2 parameters of the primary cluster of each interval on the arm. Because sums of normally distributed random variables were also normally distributed random variables, the calculation was straightforward (**Equation 3**). Using

Equation 3

$$\sum_1^I R_i \sim N(\sum_1^I \mu_i, \sum_1^I \sigma_i^2)$$

a chromosome arm was aneuploid if the following two-sided test was significant, (min(2*CDF($\sum_1^I \mu_i, \sum_1^I \sigma_i^2$), 2(1-CDF($\sum_1^I \mu_i, \sum_1^I \sigma_i^2$))) < α. To distinguish between chromosome arm gains and losses, I performed the one-sided test. For gains, 1-CDF($\sum_1^I \mu_i, \sum_1^I \sigma_i^2$) < α and for losses, CDF($\sum_1^I \mu_i, \sum_1^I \sigma_i^2$) $<$ α. For each chromosome arm, a Z-score was produced using the quantile function 1-CDF($\sum_1^I \mu_i, \sum_1^I \sigma_i^2$). Positive Z-scores represented gains and negative Z-scores represented losses.

5.5 Arm level Allelic Imbalance

Common polymorphisms from 1000G (24,720 single nucleotide and 1,500 indels, MAF>1%) were used as candidate heterozygous sites in the samples. For each of the 677 normal samples, I identified polymorphic sites that could be confidently called as heterozygous and diploid, based on their B-allele frequencies (BAF) (0.4<BAF<0.6), where BAF = #non-reference reads / total reads. BAFs were modeled at these sites as random variables from a normal distribution with mean 0.5 and a variance that depended on UID depth. The variance was estimated with maximum likelihood, as a function of UID depth (**Figure 5**). Further analysis was restricted to these sites.

To infer whether a chromosome arm in a test sample harbored allelic imbalance, I identified the subset of polymorphic sites at which both alleles were present in that sample and with sufficient UID depth (>25 UIDs). At each site, I compared the observed BAF with the normal distribution, using the expected variance for the observed UID depth, yielding a two-sided P-value. All p-values on a chromosome arm were Z-transformed and combined with a weighted Stouffer's method (**Equation 4**), with the observed UID depth at each site used as its weight.

Equation 4

$$Z \sim \frac{\sum_{i=1}^{k} w_i Z_i}{\sqrt{\sum_{i=1}^{k} w_i^2}}$$

$$w_i = UID\ Depth\ at\ Variant\ i$$

$$Z_i = Z\ score\ at\ Variant\ i$$

$$k = Number\ of\ \ Variants\ observed\ on\ a\ chromsome\ arm$$

A chromosome arm had allelic imbalance if the resulting Z score was greater than a selected statistical significance threshold.

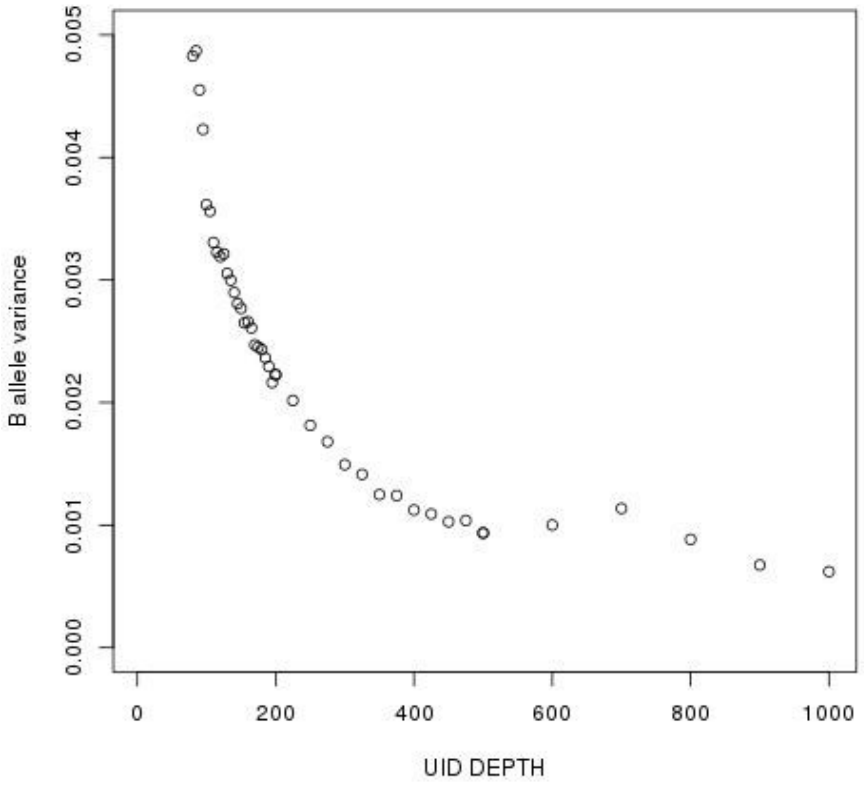

Figure 5: Empirically Estimated B allele Frequency Variance vs UID Depth.

5.6 Generalized Aneuploidy Detection

For samples with lower fractions of neoplastic cells, it may not be possible to call specific aneuploid chromosome arms with high confidence. An alternative approach was to consider all chromosome arms collectively.

A two-class support vector machine (SVM [114]) was trained to discriminate between a negative class of normal samples with no aneuploidy and a positive class of synthetic samples in which aneuploidy was spiked-in. The training set contained 677 WBC normal samples (3M-15M UIDs) and 3,150 synthetic samples. SVM training was done with the e1071 package in R, using radial basis kernel and default parameters [115]. Each sample had 34 Z-score features, representing chromosome arm gains and losses. The following arms were not included, because they were consistently miscalled as aneuploid in normal samples: 4q, 17q, 19p, 19q, 22q.

Synthetic aneuploid samples were generated by spiking in aneuploid chromosome arms to 63 of the normal WBC samples, which contained at least 9M UIDs. They were designed to represent low neoplastic cell fractions (0.005, 0.01) and degrees of aneuploidy (5,10,15,20, or 25 chromosome arm gains or losses). Each synthetic sample contained exactly 9M UIDs. Spike-ins were implemented by duplicating or subtracting UIDs from selected chromosome arms. I made a number of simplifying assumptions when generating synthetics: 1) all chromosome arms were independent; 2) chromosome arm alterations in a sample occurred in the same fraction of neoplastic cells; 3) alterations were restricted to a single copy gain or loss per arm; 4) duplication or subtraction of UIDs was restricted to UIDs that matched the reference genome.

The spike-ins were intended to represent a realistic range of arm gains and losses, based on the total numbers observed in the primary tumor samples. To ensure a wide range of alterations were included, I repeated the procedure shown in (**Figure 6**) five times.

```
N <-  number of normal WBC samples used
d <- desired degree of aneuploidy (number arms altered)
f <- desired neoplastic cell fraction
r <- arm alteration type (gain or loss)
p(r) <- probability of arm gain (0.5 default)
rho <- desired unique read depth of synthetic sample (9M default)
For i=1:N
 s <- sample[i]
 U <-   get UIDs from s
 for f in 1:F
 h <- new synthetic sample
    For d in (5,10,15,20,25)  #desired numbers of altered arms
         t <- copy of s
         For a in 1:d    #select alteration types and spike-in to t
                  r <- random select arm alteration type with p(r)
                  u_ref <- get reads mapped to a that match reference genome
                  if  r== gain
                           t  <- add  duplicate u_ref reads
                  else (r == loss)
                           t  <- subtract u_ref reads
        #serial dilution - mix spiked-in alterations to normal sample s at neoplastc cell fraction f
        n <-random select rho*f UIDs from t
        e <- random select rho * (1-f) UIDs from s
        h <- n + e
```

Figure 6: Pseudocode to Generate Synthetic Aneuploid Spike-in Samples

To balance the number of positive and negative examples used to train the SVM, the positive class was randomly subsampled to match the size of the negative class. Ten two-class SVMs were trained, by subsampling the positive class ten times, then training an SVM on the subsampled positive class and the negative class. Each sample to be classified was scored by all ten SVMs, and the ten scores were averaged to yield a final score.

Samples may have a wide range of UID depths, which I observed to confound SVM scoring, so that samples with very low UID depth received artificially high scores. I controlled for UID depth by modeling the change in SVM scores as a function of UID depth in normal samples. Each of the 63 WBC normal samples was randomly down-sampled to yield ten replicate synthetic samples of UID depth ranging from 100K to 9M. Gain/loss Z-scores were computed for all chromosome arms and each synthetic sample was scored by ten SVMs and the scores were averaged, using the protocol previously described. This procedure yielded 630 SVM scores for synthetic normal samples at each UID depth. All scores were converted to ratios by finding the sample at each UID depth with the minimum SVM score and dividing all scores at the same depth by that value. The average ratio *r* at each depth decreased monotonically as a function of increasing UID depth **Figure 7**). The relation between UID depth and SVM score was model using **Equation 5** (A=-7.076*10^-7 and B=-1.946*10^-1) and raw SVM scores were corrected by dividing by ratio *r*.

Equation 5

$$\log\left(1-\frac{1}{r}\right) = Ax + B$$

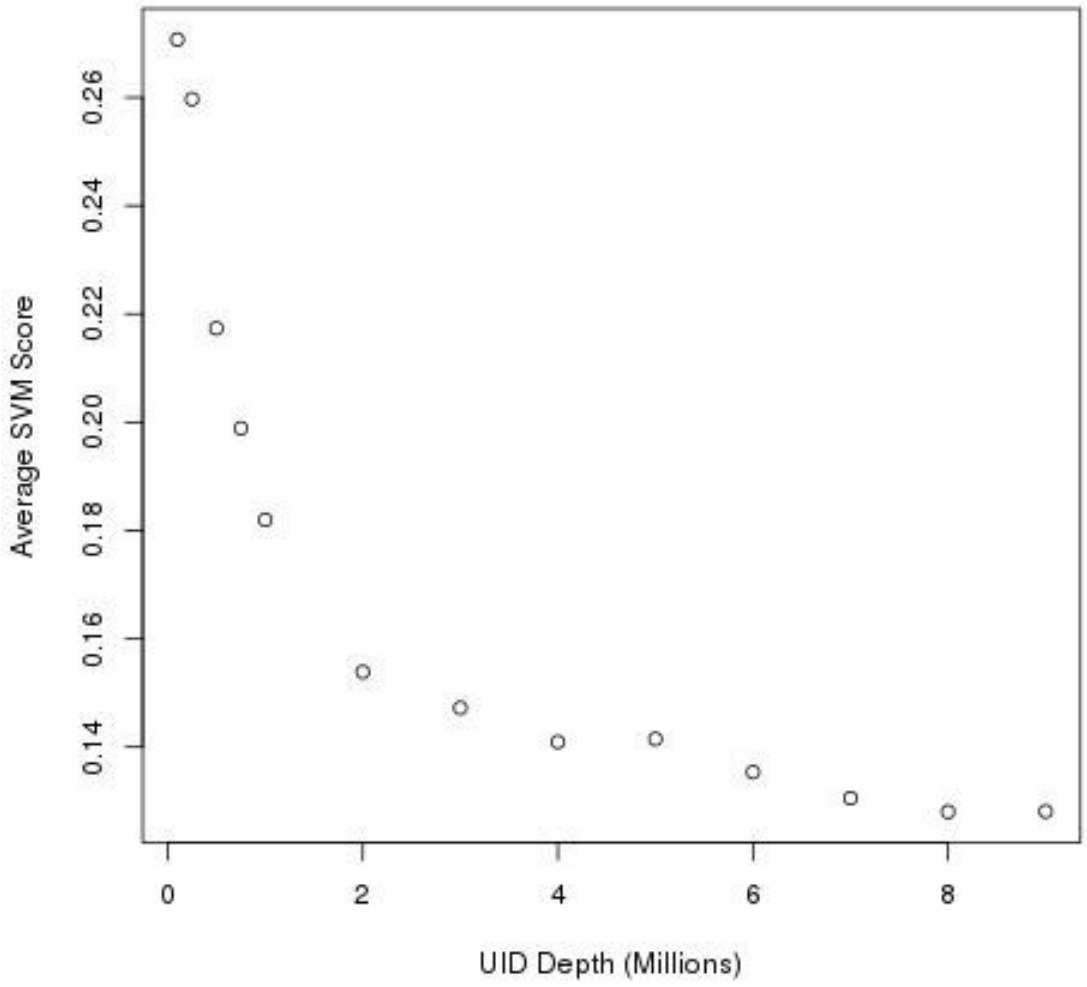

Figure 7: Raw SVM Scores as function of UID Depth

Finally, because the SVM was designed for difficult-to-detect aneuploidies, before classifying a sample, I checked whether it contained at least one aneuploid chromosome arm that could be identified with high confidence. If such a chromosome arm was present, the sample was called as aneuploid and SVM classification was not applied. A "high confidence" call was made if a chromosome received a Z-score (gain, loss, or allelic imbalance) that was an extreme outlier ($\pm 4\ \sigma$ from max or min Z-score) with respect to the distribution of Z-scores seen in the 677 normal WBC samples.

5.7 Somatic sequence mutations and microsatellite instability (MSI)

When matched normal samples were available, somatic single nucleotide substitution (SBS), insertion and deletion (indel) mutations based on FastSeqs LINE amplicon sequences and alignments could be identified. The SBS mutations were identified by directly comparing amplicons from the test sample with amplicons from the matched normal, and did not require any alignment to the reference genome. I used the following conservative thresholds. Amplicons with fewer than 200 reads and 50 UIDs in the matched normal were excluded. Amplicons that occurred only in the test sample were identified, and any of these amplicons with fewer than 20 reads and 5 UIDs were excluded. The remaining reads were compared position-wise to the reads from the normal sample to identify test sample reads that differed from any normal read by exactly one nucleotide substitution. This procedure enabled us to call SBS somatic mutations and subsequently estimate mutation load and the presence of known carcinogen signatures.

Somatic indel mutations were called by aligning amplicons from the test sample and matched normal sample to the reference genome (GRc37) with Bowtie2 [116]. To reduce potentially including PCR artifacts, amplicons with a ratio of reads to UIDs less than two were excluded. If Bowtie2 reported an insertion or deletion in the test sample but not in the matched normal, and the amplicon had >= 10 UIDs in the test sample and in the matched normal, it was considered to harbor a somatic indel then detect the number of somatic indels in monotracts of > 3 nucleotides. There were 17,488 of these monotracts in the FastSeqs amplicons. The number of somatic monotract indels in a normal sample can be modeled using a Poisson distribution where lambda is the average number of somatic indels resulting of a sequencing artifact. Samples with statistically significant quatities of somatic indels in monotracts can be identified as MSI.

5.8 Sample Identification

This protocol can be used to estimate concordance between polymorphic sites in two samples of interest. Samples were sequenced and all amplicons were aligned to reference genome GRC37 with Bowtie2. Amplicons with <10 UIDs in either sample were excluded. The 1000G common polymorphisms (Sample alignment and genomic interval grouping) were used to identify the genotypes at 26,220 sites in each sample. Each polymorphic site was called as "0" (homozygous reference, >0.95 UIDs matching reference allele), "1" (heterozygous, 0.05-0.95 UIDS matching either reference or alternate allele, >10 UIDs for each allele), or "2" (homozyous alternate, >0.95 UIDs matching alternate allele). A broad UID range was used to call heterozygous sites to ensure that heterozygosity could be detected in tumor samples harboring allelic imbalance. To compare two samples, I considered polymorphic sites with sufficient

amplicon coverage in both samples and counted the number of sites that agreed on “0”, “1”, and “2”. Concordance was the number of matched polymorphic sites divided by the total number of covered polymorphic sites

Chapter 6: WALDO Performance and Applications

First, I compared this protocol to whole genome sequencing and then performance was assessed on the ability to detect: 1) single chromosome arm events using synthetic data; 2) identifying chromosome arm aneuploidies in primary tumors; 3) detecting general aneuploidy using synthetic data; and 4) Mutation load, carcinogenic signatures, and MSI.

6.1 Comparison to Whole Genome Sequencing

Amplifying and sequencing a discrete set of locations using a single primer has numerous experimental and computational advantages over whole genome sequencing. Here, I directly compare the performance of WALDO to whole genome sequencing. To my knowledge, WISECONDOR is the only free publically available prenatal screening protocol for whole genome sequencing [117]. WISECONDOR partitions the genome into intervals. The genomic intervals are grouped together based on similar properties. These groups of intervals are used to perform statistical testing. Chromosomal aneuploidies can be detected using a sliding window approach that identifies intervals along a chromosome that have been altered. According to the authors of WISECONDOR, cases and controls should always come be performed on the same sequencing machine, aligned using BWA [118], sorted using the picard software (http://broadinstitute.github.io/picard), and filtered using samtools [119]. Default parameters of WISECONDOR were used and Trisomy 21 was called using the "windowed, bin test."

In order to directly compare these approaches, 12 mixtures were created using 2 ng of normal DNA and 0.2 ng of trisomy 21 DNA. The mixtures were created to replicate typical fetal fractions in noninvasive prenatal testing (approximately 10%). Trisomy 21 and normal samples were sequenced and analyzed using both FAST-SeqS + WALDO and whole genome sequencing + WISECONDOR. Sensitivity (ability to correctly identify Trisomy 21) and specificity (ability to correctly identify normal samples as normal) were calculated at different read depths. WALDO identifies chromosome gains and losses at different thresholds depending on the desired alpha where as WISECONDOR calls chromosome gains and losses depending on the number of altered genomic intervals. WISECONDOR does not provide a way to estimate the type I error rate and the authors report no false positives for chromosome 21. While both methods can reliably detect trisomy 21 at high specificities required for NIPT, WALDO and FAST-SeqS outperforms whole genome sequencing at lower read depths (**Figure 8).**

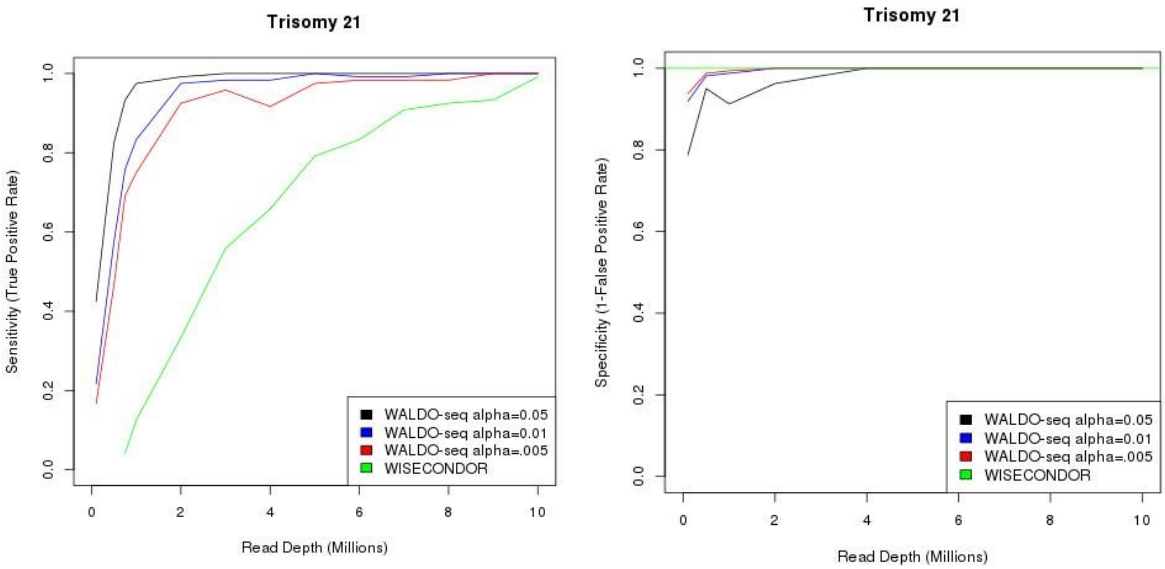

Figure 8: Comparison to Whole Genome Sequencing

6.2 Detection of Single Chromosome Arm Events Using Synthetic Data

I generated synthetic aneuploid samples derived from 63 normal WBC samples. 677 normal WBC samples and the synthetic samples were used to evaluate performance on single chromosome arm events (gains/losses and allelic imbalance). For each of the 63 samples, gains and losses were generated for each chromosome arm at various cell fractions. WALDO provides a statistical framework to assess the significance of each chromosome arm. Chromosome arm gains/losses and allelic imbalance can be identified at the desired alpha (type I error rate). I assessed performance for all chromosome arms (**Figure 9**). Sensitivity is the ability to correctly identify a chromosome arm (gain/loss or allelic imbalance) at a given statistical alpha. Specificity is the ability to correctly identify normal chromosome arms as normal for a given alpha. In theory, specificity is equal to 1 minus alpha (1- the type I error rate). For gains/losses across all chromosomes the specificities for alphas 0.05, 0.01, and 0.005 were 0.969, 0.995, and 0.998 respectively. The observed statistical significances were slightly more conservative than expected for a given alpha. For allelic imbalance across all chromosomes the specificities for alphas 0.05, 0.01, and 0.005 were 0.946, 0.970, and 0.992 respectively. The observed statistical significances were slightly less conservative than expected. Gain/losses sensitivity approaches 100% for cell fractions above 5%. Sensitivity can be improved depending on the application and acceptable type I error rate. Allelic imbalance, however, was a much weaker signal and not informative for cell fractions below 5%. Performance varied depending on the specific chromosome arm and unsurprisingly larger chromosomes were much easier to detect gains/losses. NIPT testing typically evaluates trisomies 13, 18, and

21. Fetal cell fractions typically observed in NIPT are approximately 10% and testing is often not performed on cell fractions < 5% [117]. At 10 % cell fraction using an alpha of 0.005, WALDO had sensitivities 100%,100%, 100% and specificities 99.7%, 100%, 99.4% for 13q,18q,21q respectively. For clinical use, thresholds should be optimized for each chromosome arm to optimize the tradeoffs of sensitivity and specificity.

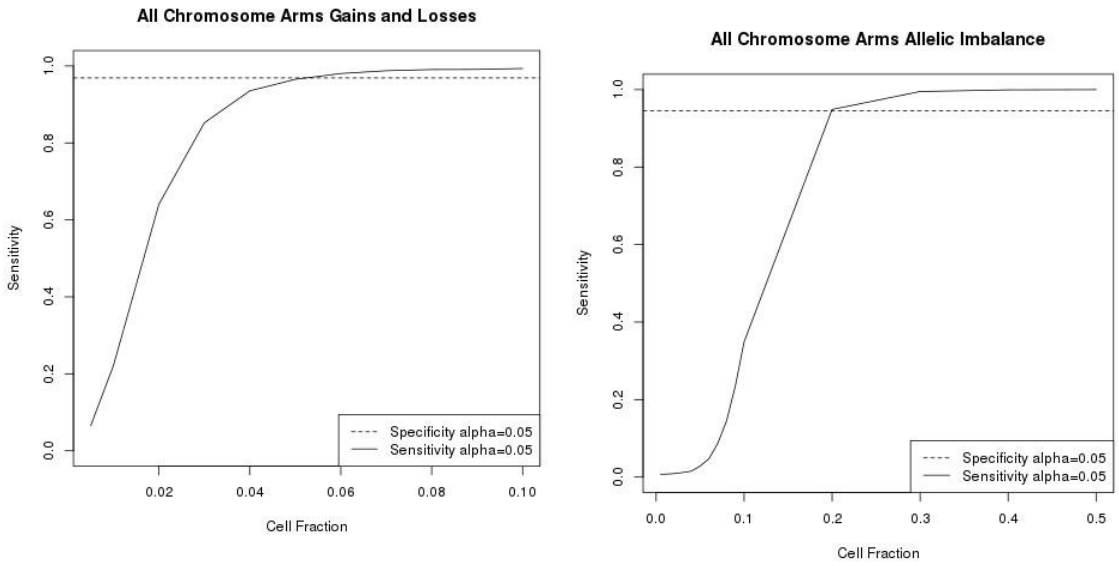

Figure 9: Chromosome Arm Aneuploidy Detection at alpha = 0.05

6.2 Identifying Chromosome Arm Aneuploidies in Tumors

I applied the framework to identify single chromosome arm gains and losses in tumor samples. I analyzed aneuploidy in 546 primary tumors that consisted of 45 bladder urothelial carcinomas (BRCA), 56 breast invasive carcinomas (BRCA), 302 colon and colorectal adenocarcinomas (COAD and COADREAD) cancers, 26 esophageal carcinomas (ESCA), 37 head and neck squamous cell carcinomas (HNSC), 27 liver cancers (LIHC), 9 ovarian cancers (OV), 22 uterine corpus endometrial carcinomas (UCEC), 22 stomach adenocarcinomas (STAD), and 32 colorectal adenomas. I identified gains and losses using a z threshold of 3 and -3 and quantified the total numbers of aneuploidies in each sample. I compared the WALDO aneuploidies to the aneuploidies detected from array CGH in The Cancer Genome Atlas (TCGA) [120] (**Table 5**). While I do not expect the analyses to match exactly because different samples were being used, the patters in specific cancer types as well as across cancer types should be very similar. On average WALDO identified more aneuploidies than the TCGA analysis (15.6 vs 13.0) but the distribution of altered chromosome arms was very similar (**Figure 10**) suggesting that WALDO can chromosomal events in actual cancer tumors.

Table 5: Aneuploidy Analysis of Tumors

	WALDO Total Aneuploidies	TCGA Aneuploidies	WALDO Gains	TCGA Gains	WALDO Losses	TCGA Losses
ALL	15.6	13.0	7.9	6.2	7.7	6.7
BRCA	16.5	13.0	8.2	6.0	8.3	7.0
BLCA	16.4	13.5	9.0	6.9	7.4	6.7
COAD COADREAD	15.0	12.3	7.1	6.7	7.9	5.6
ESCA	18.7	16.0	9.7	7.8	8.9	8.2
HNSC	15.6	11.1	8.5	5.9	7.1	5.2
LIHC	15.6	11.6	8.1	5.4	7.5	6.2
OV	20.4	19.3	10.0	7.4	10.4	11.8
UCEC	5.9	7.3	3.4	3.6	2.5	3.7
STAD	16.9	10.7	9.1	5.6	7.8	5.0
Colorectal Adenomas	8.3	N/A	5.0	6.2	N/A	6.7

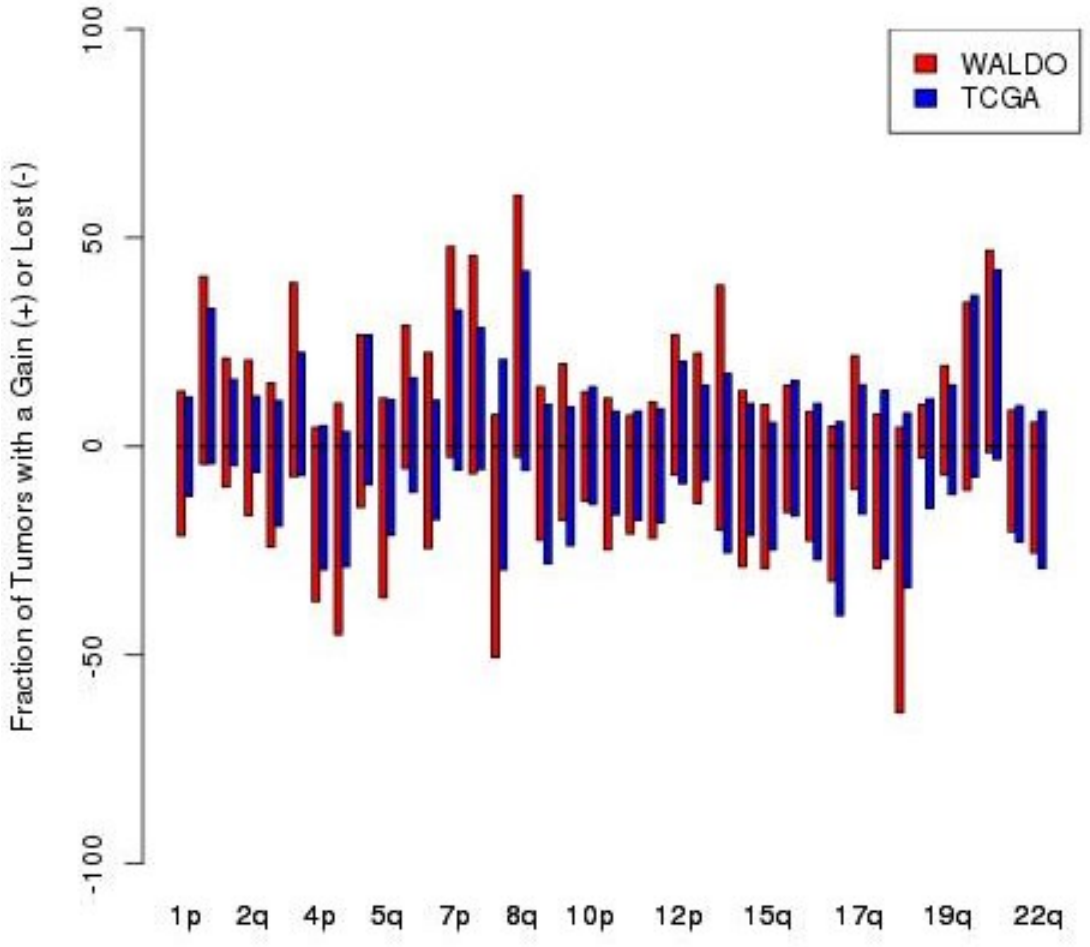

Figure 10: Distribution of Aneuploid Chromosomes Arms across All Cancer Types

6.3 Detection of General Aneuploidy Using Synthetic Data

Aneuploidy is a general feature of neoplastic cells [121]. I investigated whether generalized aneuploidy prediction (identifying the presence of any aneuploid chromosomes) could discriminate between synthetic aneuploid samples and normal samples. Synthetic samples were created with multiple aneuploidies (5,10,15,20,25) at various cell fractions. WALDO summarizes the chromosome arm Z scores using a supervised machine learning method. To evaluate the discriminatory power of WALDO generalized aneuploidy score, I partitioned the training set into 11 equal folds (11 folds of 16 WBC samples). Technical and synthetic replicates were restricted to the same fold. WALDO detect aneuploidy in reliable detect aneuploidy at close to 100% of the time in samples with more than 1% neoplastic content at very high levels of specificity (99%). Samples with low neoplastic content (0.5%, 1%) can also be detected at high levels of specificity (99%) when there is a large number of gains and losses present in the sample (**Figure 11**). Performance was measure at 9M UIDs, however, increasing the number of UIDs in sequencing can improve sensitivity without decreasing specificity (**Figure 12**).

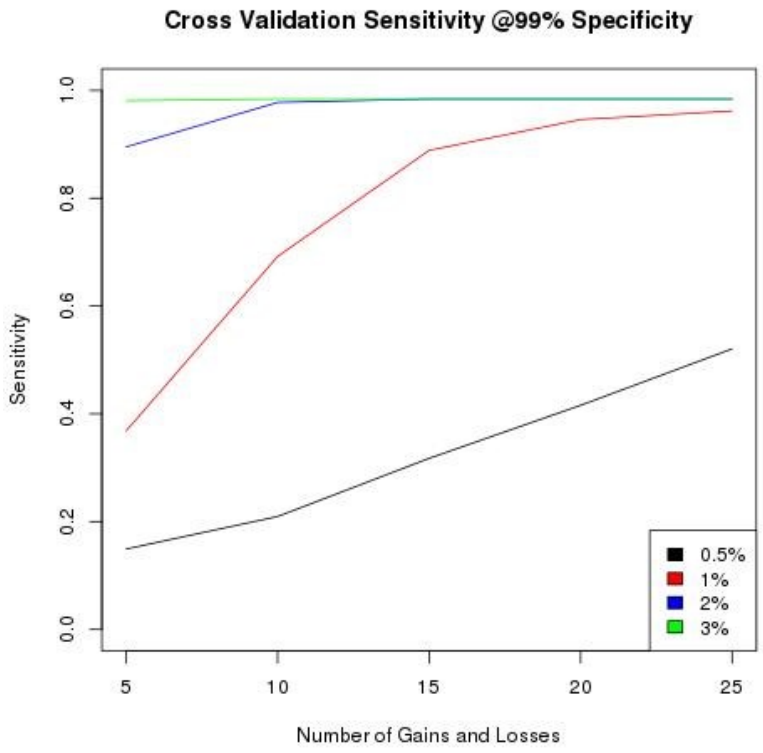

Figure 11: Cross Validation Performance of Generalized Aneuploidy Detection on Sythetic Aneuploid Samples

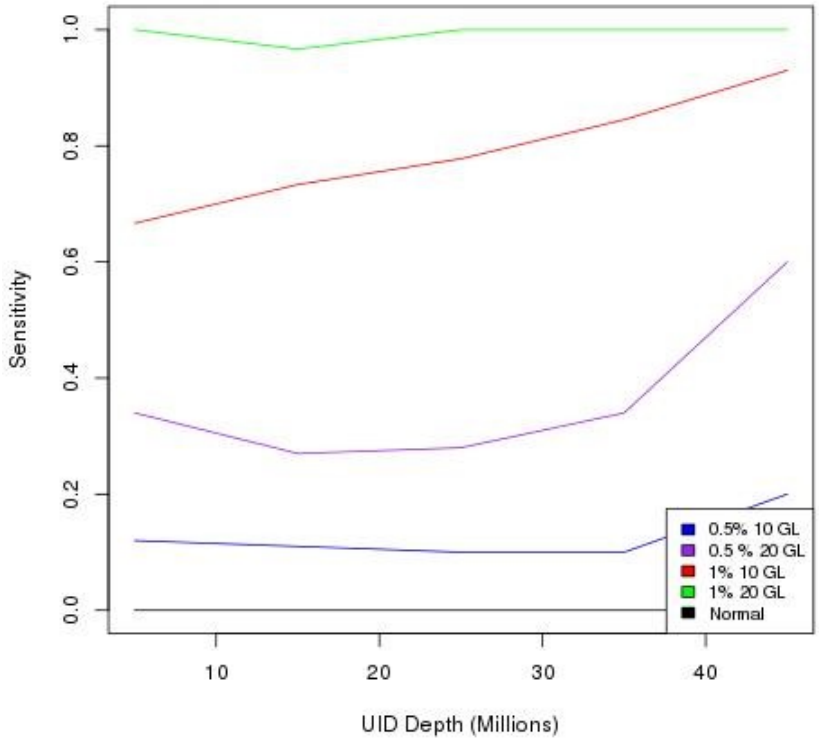

Figure 12: Generalized Aneuploidy Performance on High UID Depth Samples

6.4 Mutation load, carcinogenic signatures, MSI

To my knowledge, amplicon sequencing has not previously been used to estimate somatic mutation load, or to identify carcinogenic signatures or microsatellite instability (MSI) in primary tumor samples. Whole-exome or whole-genome sequencing protocols are typically chosen. This obtained good estimates of mutation load in 10 urothelial carcinomas of the upper urinary tract (UTUCs), to identify an aristolochic acid mutation signature in 6 of these UTUCs, and to identify MSI in 6 mismatch-repair (MMR) deficient colorectal cancers (CRCs). This required matched normal samples to call somatic single nucleotide substitution (SBS) mutations and indels. I applied Fast-SeqS and Safe-SeqS amplicon sequencing and the somatic mutation calling protocol. For each sample, the total number of somatic SBS mutations was counted (mutation load) (**Figure 13**) and the fraction of each possible nucleotide substitution was determined (A->T, A->C, etc.) (mutation spectrum) . Sample mutation load and mutation spectrum were highly correlated with estimates based on previous whole exome sequencing [122]. (R^2=0.98 and 0.95, respectively). The mutation spectrum noticeably contained the aristolochic acid mutation signature (A->T, T->A) (**Figure 14**).

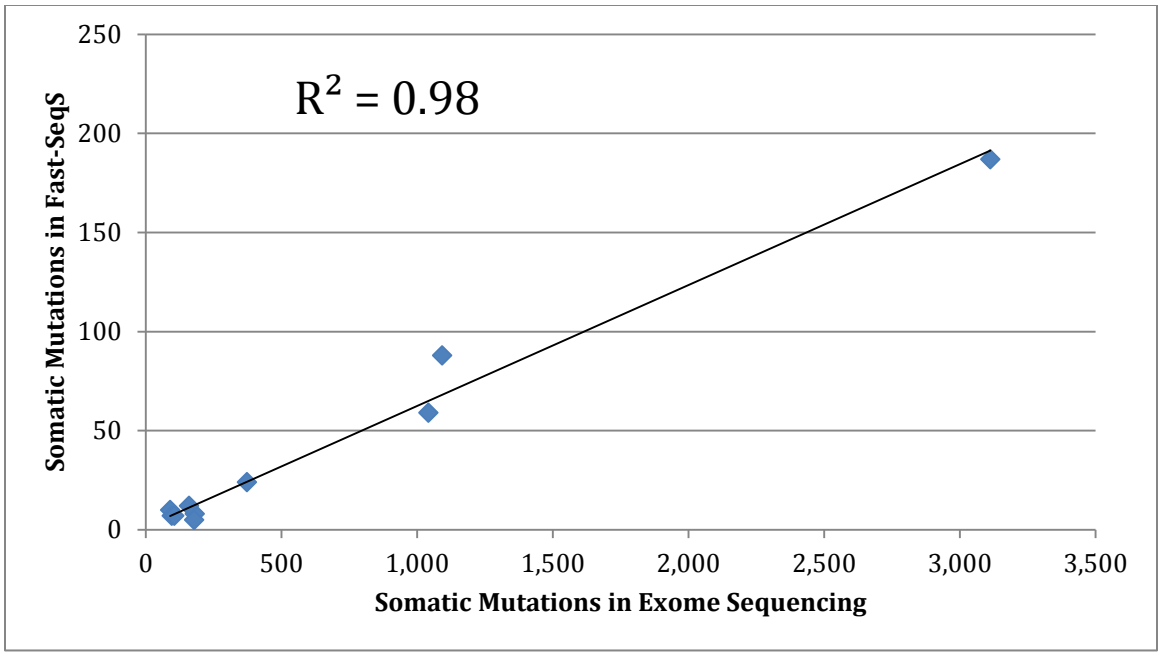

Figure 13: Mutation Load Comparison to Exome Sequencing

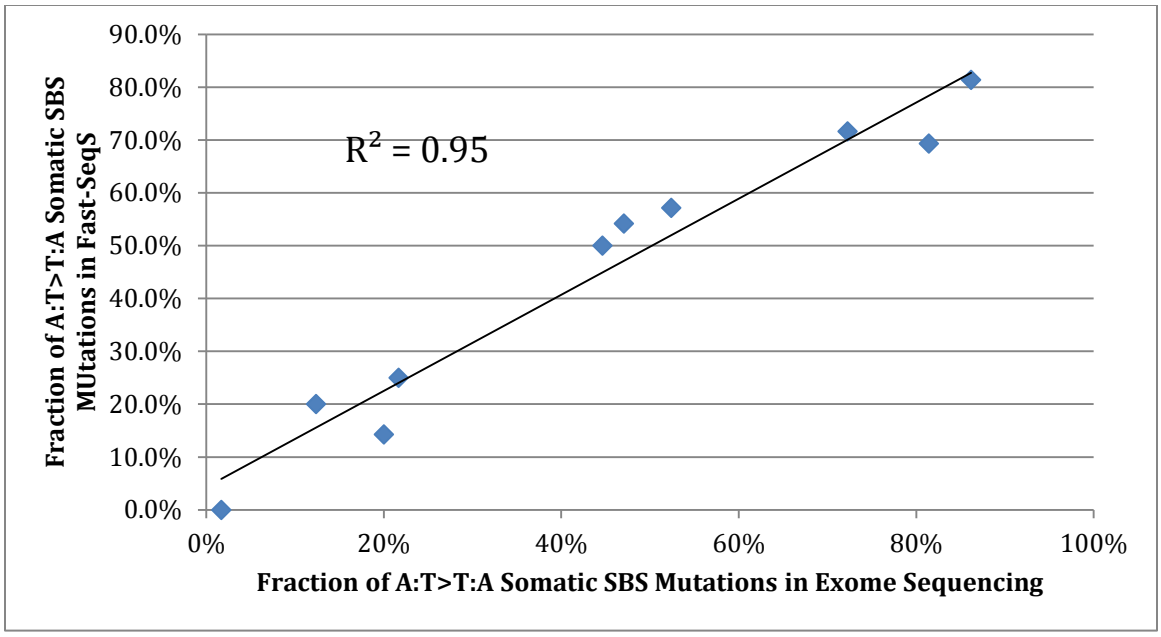

Figure 14: Mutation Spectrum Comparison to Exome Sequencing

Furthermore, the FastSeqs amplicons harbored 17,488 monotracts of >3 nucleotides, which enabled identification of MSI, by counting the number of somatic indels that occurred in the monotracts and identifying statistically significant samples (See 5.7 Somatic sequence mutations and microsatellite instability (MSI)). I applied this approach to the 6 MSI colorectal tumors and 10 UTUC tumors (**Table 6**). As a negative control, the 16 matched normal samples were partitioned into equal parts and somatic indels were called. No normal samples were statistically significant. All MSI tumors were statistically significant. SB 102 PT was the one statistically significant UTUC tumor. SB 102 PT had 3,112 somatic mutations identified in exome sequencing [122]. The other UTUC tumors had ~100 somatic mutations identified during exome sequencing so it was unsurprising that this sample was statistically significant using the protocol to detect MSI.

Table 6: MSI Sample Data

Sample Name	Sample Type	Somatic Indels Disrupting Monotracts	P-value
Co 083	MSI	27	1.25E-30
Co 083 N	Normal	0	0.632
Co 086	MSI	29	1.43E-33
Co 086 N	Normal	1	0.264
Co 088	MSI	30	4.62E-35
Co 088 N	Normal	0	0.632
Cx 002	MSI	10	1.00E-08
Cx 002 N	Normal	0	0.632
Cx 010	MSI	45	6.83E-59
Cx 010 N	Normal	0	0.632
Hx 075	MSI	67	1.51E-97
Hx 075 N	Normal	0	0.632
SB 101 N	Normal	0	0.632
SB 101 PT	UTUC	1	0.264
SB 102 N	Normal	0	0.632
SB 102 PT	UTUC	20	7.54E-21
SB 103 N	Normal	0	0.632
SB 103 PT	UTUC	2	0.0803
SB 104 N	Normal	1	0.264
SB 104 PT	UTUC	0	0.632
SB 105 N	Normal	0	0.632
SB 105 PT	UTUTC	1	0.264
SB 106 N	Normal	0	0.632
SB 106 PT	UTUC	0	0.632
SB 111 N	Normal	1	0.264
SB 111 PT	UTUC	0	0.632
SB 112 N	Normal	0	0.632
SB 112 PT	UTUC	0	0.632
SB 113 N	Normal	0	0.632
SB 113 PT	UTUC	0	0.632
SB 114 N	Normal	0	0.632
SB 114 PT	UTUC	0	0.632
SB 115 N	Normal	0	0.632

SB 115 PT	UTUC	1	0.264
SB 116 N	Normal	0	0.632
SB 116 PT	UTUC	0	0.632

6.5 Applications and Discussion

Aneuploidy is a feature of most cancer cells, and cancer malignancy is related to the amount of aneuploidy. In the past decade, many techniques have been developed to detect aneuploidy based on whole genome and targeted sequencing protocols. Here, I introduce a new analytical approach for aneuploidy detection from FastSeqs amplicon sequencing of long interspersed nucleotide elements (LINEs). FastSeqs is fast and efficient but yields a read depth distribution that is not well handled by computational methods developed for WGS and targeted sequencing. This method uses a within-sample approach to identify chromosome arm level gains and losses and a machine learning method to summarize the general aneuploidy of a sample. In collaboration with the Johns Hopkins Ludwig Cancer Center, this method is currently being applied to numerous ongoing studies. This method is being used to detect aneuploidy in pancreatic cysts, urine, stool, plasma, cerebral spinal fluid, pap smears, and saliva.

Chapter 7: Concluding Remarks and Future Work

Advances in sequencing technology have greatly reduced the costs incurred in collecting raw sequencing data. Academic laboratories and researchers therefore now have access to very large datasets of genomic alterations but limited time and computational resources to analyze their potential biological importance. Computational tools can assess the potential impact genetic variation has upon human health. These tools can narrow down long lists of variants to identify a small subset of candidate variants. Despite the utility of computational tools, most can only predict the impact of single nucleotide changes. My dissertation introduced two novel methods that can be applied to genetic variation beyond single nucleotide changes.

7.1 VEST-indel

In the first part of my thesis, I develop a method for predicting indel pathogenicity. VEST-indel differs from current indel classifiers that were developed to predict the impact of indels on protein structure or function. VEST-indel has substantial improvement in specificity relative to existing methods, highlighting reductions in falsely classifying neutral variants as pathogenic largely from the incorporation of a new text mining feature that captures the known relevance of a gene to human health. VEST-indel predicts pathogenicity in both inframe and frameshift indel and can be used in combination with the original VEST missense method to perform joint prioritization. This enables researchers to perform single pass sorting on all the sequence variants detected in their sequencing studies.

I demonstrated the utility of combining in-frame, frameshift, and missense variants for reliable joint prioritization. In the future, I would like to see the Variant Effect Scoring Tool extended beyond these three variant types. I think there is a pressing need to create pathogenicity predictors specifically designed for splice variants, regulatory variants, as well as variants that disrupt or introduce stop codons. Each of these predictors could be combined using the same statistical framework proposed in my thesis. Currently, the CADD method already provides joint prioritization across all of these variant types. The CADD method, however, was designed very generally and not for any particular variant type. In my thesis, I demonstrated, the method has significant limitations. I believe that developing individual predictors for each variant type and then combining the scores using a statistical framework can greatly improve single pass prioritization across all types of genetic variation.

7.2 WALDO

Aneuploidy or abnormal numbers of chromosomes, is an important feature in both prenatal screening and cancer detection. I introduce a new analytical approach for aneuploidy detection from FastSeqs amplicon sequencing of long interspersed nucleotide elements (LINEs). FastSeqs is fast and efficient but yields a read depth distribution that is not well handled by computational methods developed for WGS and targeted sequencing. This method uses a within-sample approach to identify chromosome arm level gains and losses and a machine learning method to summarize the general aneuploidy of a sample. It can also be used to identify mutation load, carcinogen signatures, and microsatellite instability. It is effective on samples containing only a few ng of DNA and as little as 1% neoplastic content.

I believe aneuploidy detection can be improved with the presence of a matched normal sample. With a matched normal, I believe WALDO could detect aneuploidy in much less than 1% neoplastic content. FastSeqS amplicons are highly repetitive. Private non-reference alleles can cause mis-mapped amplicons during alignment. These mis-mapped amplicons appear as amplifications or deletions in a test sample. WALDO was designed to achieve a high specificity and treated potential mis-mapping as noise. With a matched normals, private variants could be detected thus improving alignment and ultimately reducing noise when identifying aneuploidy.

7.3 Conclusion

High-throughput sequencing technology routinely generates millions of genetic variants. Computational algorithms can reduce the time required to analyze the potential impact on disease etiology of the many genetic variants detected and there is a pressing need for tools that can handle variation beyond single nucleotide changes. In this dissertation, I introduced two novel computational methods (VEST-indel and WALDO) for pathogenicity prediction of indels, aneuploidy detection, and microsatellite detection from somatic indels.

Appendix A: Supplementary Tables and Figures

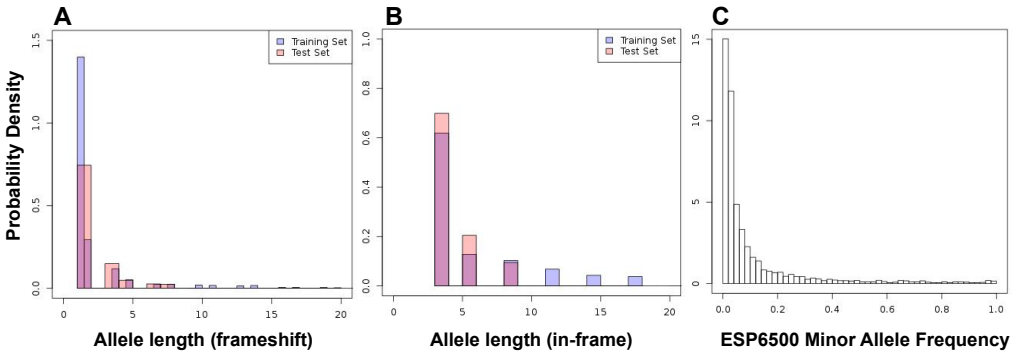

Figure S 1: Histograms for allele length and minor allele frequency for selected data sets. Indel length probability density for both frameshift (A) and in-frame (B) indels, separated by training and testing datasets. Figure C plots the minor allele frequency (MAF) probability density for ESP6500 indels used in the benign training set. Pathogenic variants from HGMD and ClinVar, and benign interspecies indels, do not have reported MAFs and are therefore absent from Figure S1.

Table S 1: All Features Considered During VEST-indel Development:

Feature Categories	Features	Acronym	In-frame importance	In-frame rank	Frameshift importance	Frameshift rank
DNA sequence conservation	Exon Conservation	ExonConservation	10.76	4	10.8	3
	Hidden Markov model score of alanine substitution at first position where indel occurs from 46-way mammalian genome alignments	AluSubPHC	10.67	5	11.56	2
DNA natural variation	ExonSnpDensity	ExonSnpDensity	6.68	8	7.48	5
Gene-level annotations	Log10 of count of publications in PubMed in which gene in which variant occurs is named in title or abstract	PubMed	24.18	1	26.41	1
	Number of codons (in CDS) downstream of the codon where the insertion or deletion begins.	C-terminal Remainder	3.43	17	4.42	8
	For insertions, this length is always 0. For deletions, this length is the number of deleted nucleotides.	Length of Reference Allele	4.19	14	2.44	13
	For insertions, this length is the number of inserted nucleotides. For deletions, this length is always 0.	Length of Alternate Allele	0.75	29	2.51	11
Transcript-level annotations	The number of transcripts in the SNVBox database (from RefSeq and Ensembl) that cover the codon where the insertion or deletion begins	Transcripts	1.81	22	6.04	6
Computational predictions of protein local structure	High temperature factor	PredBFactorF	10.34	6	2.11	17
	Low stability	PredStabilityL	4.97	12	2.25	15
	Solvent accessibility	PredRSAE	6.49	10	2.03	18

Feature Categories	Features	Acronym	In-frame importance	In-frame rank	Frameshift importance	Frameshift rank
	helical secondary structure	PredSSH	1.71	23	-0.56	29
Protein local regional sequence composition	Postive formal charge	RegCompKR	3.76	16	-1.1	33
	Glutamine enriched	RegCompQ	3.28	18	-0.19	25
	Proline enriched	RegCompP	3.98	15	4.84	7
	Cysteine enriched	RegCompC	4.48	13	-0.18	24
	Glycine enriched	RegCompG	3.01	19	1.45	20
	Histidine enriched	RegCompH	-1.27	33	-0.42	28
	Hydrophobic	RegCompILVM	6.55	9	2.62	10
	Aromatic	RegCompWYF	11.57	3	2.29	14
	Negative formal charge	RegCompDE	5.5	11	2.81	9
	Shannon entropy of amino acids surrounding position	RegCompEntropy	14.28	2	7.94	4
Protein-level annotations from UniProt	Selenocysteine	UniprotSECYS	N/A	N/A	N/A	N/A
	Zinc finger	UniprotZNFINGER	0.81	28	-0.84	30
	RNA-binding domain	UniprotDOM_RNABD	N/A	N/A	N/A	N/A
	DNA-binding domain	UniprotDNABIND	1.43	25	N/A	N/A
	in a protein domain		N/A	N/A	N/A	N/A
	calcium-binding site	UniprotCABIND	N/A	N/A	N/A	N/A
	post-translational modification site	UniprotDOM_PostModRec	N/A	N/A	N/A	N/A
	transcription factor domain	UniprotDOM_TF	N/A	N/A	N/A	N/A
	any regional annotation identified	UniprotREGIONS	1.28	26	1.92	19
	pro-peptide region	UniprotPROPEP	1.01	27	-0.29	27
	site in a domain that binds to the cell membrane	UniprotDOM_MMBRBD	N/A	N/A	-1.43	35
	compositionally biased region	UniprotCOMPBIAS	6.96	7	2.48	12
	any site annotation identified	UniprotMODRES	N/A	N/A	N/A	N/A
	binding site of any kind	UniprotBINDING	N/A	N/A	N/A	N/A
	disulfide site	UniprotDISULFID	N/A	N/A	N/A	N/A

Feature Categories	Features	Acronym	In-frame importance	In-frame rank	Frameshift importance	Frameshift rank
	site of a domain that determines correct cellular localization	UniprotDOM_LOC	-0.01	32	-1.28	34
	site of localization signal (protein targeted to secretory pathway or periplasm)	UniprotSIGNAL	2.69	20	1.13	21
	transmembrane domain	UniprotTRANSMEM	1.71	23	-0.19	25
	site in a domain involved in chromatin structure remodeling	UniprotDOM_Chrom	N/A	N/A	N/A	N/A
	nucleotide phosphate-binding restion	UniprotNPBIND	N/A	N/A	1.01	22
	site in an enzymatic domain responsible for any kind of post-translational modification	UniprotDOM_PostModEnz	0.44	31	-0.17	23
	repeat region	UniprotREP	2.58	21	-0.93	31
	metal-binding site	UniprotMETAL	N/A	N/A	N/A	N/A
	site of a known functional motif	UniprotMOTIF	N/A	N/A	N/A	N/A
	site in a protein-protein interaction domain	UniprotDOM_PPI	0.67	30	2.24	16
	lipid-binding site	UniprotLIMID	N/A	N/A	N/A	N/A
	site involved in enzymatic activity	UniprotACTSITE	N/A	N/A	N/A	N/A
	carbohydrate-binding site	UniprotCARBOHYD	N/A	N/A	-1.01	32

Candidate features are divided into seven categories: DNA sequence conservation, DNA natural variation, gene-level annotations, transcript-level annotations, computational predictions of protein local structure, protein local regional sequence composition, amino acid residue conservation, and protein-level annotations from UniProtKB. Each feature is described and labeled by the acronym used in the manuscript. Feature importance Z-score for in-frame and out-of-frame classifiers shown in last two columns. N/A = The feature was not selected as important by the Random Forest.

Table S 2: **VEST-indel Feature Importance Ranking**

Features Number	In-frame Feature	AUC ROC	Included?	Frameshift Feature	AUC ROC	Included?
1	PubMed	0.761	Yes	PubMed	0.806	Yes
2	RegCompEntropy	0.905	Yes	AluSubPHC	0.872	Yes
3	RegCompWYF	0.916	Yes	ExonConservation	0.882	Yes
4	ExonConservation	0.931	Yes	RegCompEntropy	0.897	Yes
5	AluSubPHC	0.937	Yes	ExonSnpDensity	0.905	Yes
6	PredBFactorF	0.948	Yes	Transcripts	0.907	Yes
7	UniprotCOMPBIAS	0.944	Yes	RegCompP	0.901	Yes
8	ExonSnpDensity	0.948	Yes	C-terminal Remainder	0.905	Yes
9	RegCompILVM	0.947	Yes	RegCompDE	0.904	Yes
10	PredRSAE	0.948	Yes	RegCompILVM	0.906	Yes
11	RegCompDE	0.950	Yes	Length of Alternate Allele	0.911	Yes
12	PredStabilityL	0.946	Yes	UniprotCOMPBIAS	0.907	Yes
13	RegCompC	0.949	Yes	Length of Reference Allele	0.915	Yes
14	Length of Reference Allele	0.948	Yes	RegCompWYF	0.911	Yes
15	RegCompP	0.946	Yes	PredStabilityL	0.915	Yes
16	RegCompKR	0.948	Yes	UniprotDOM_PPI	0.917	Yes
17	C-terminal Remainder	0.947	Yes	PredBFactorF	0.915	No
18	RegCompQ	0.948	Yes	PredRSAE	0.913	No
19	RegCompG	0.951	Yes	UniprotREGIONS	0.914	No
20	UniprotSIGNAL	0.950	Yes	RegCompG	0.914	No
21	UniprotREP	0.952	Yes	UniprotSIGNAL	0.914	No
22	Transcripts	0.949	Yes	UniprotNPBIND	0.914	No
23	PredSSH	0.952	Yes	UniprotDOM_PostModEnz	0.914	No
24	UniprotTRANSMEM	0.948	No	RegCompC	0.914	No
25	UniprotDNABIND	0.948	No	RegCompQ	0.914	No
26	UniprotREGIONS	0.948	No	UniprotTRANSMEM	0.914	No
27	UniprotPROPEP	0.948	No	UniprotPROPEP	0.914	No
28	UniprotZNFINGER	0.948	No	RegCompH	0.914	No
29	Length of Alternate Allele	0.948	No	PredSSH	0.914	No
30	UniprotDOM_PPI	0.948	No	UniprotZNFINGER	0.914	No

AUC was calculated after the addition of the listed feature using an independent feature selection set of Pathogenic and Benign Variants.

Table S 3: **Matrix Comparing features of each method**

Feature Categories	**VEST-indel In-frame**	**Sift In-frame**	**Ddig-in In-frame**	**VEST-indel Frameshift**	**Sift Frameshift**	**Ddig-in Frameshift**	**CADD**
DNA sequence conservation	Exon Conservation score	The conservation score of the DNA base to the left of the allele	Average, Maximum DNA conservation score	ExonConservation score	Fraction of affected conserved DNA bases	window-based DNA conservation score, Ka/Ks ratio	PhastCons scores, PhyloP scores, Gerp scores
	Hidden Markov model score of alanine substitution at first position where indel occurs from 46-way mammalian genome alignments			Hidden Markov model score of alanine substitution at first position where indel occurs from 46-way mammalian genome alignments			
DNA natural variation	ExonSnpDensity			ExonSnpDensity			
Gene-level annotations	Log10 of count of publications in PubMed in which gene in which variant occurs is named in title or abstract			Log10 of count of publications in PubMed in which gene in which variant occurs is named in title or abstract			
	Insertion Length		Insertion Length	Insertion Length			Insertion Length, Reference Allele
	Deletion Length		Deletion Length	Deletion Length			Deletion Length, Alternate Allele
Transcript-level annotations	The number of transcripts in the SNVBox database (from RefSeq and Ensembl) that cover the codon where the insertion or deletion begins			The number of transcripts in the SNVBox database (from RefSeq and Ensembl) that cover the codon where the insertion or deletion begins		Minimum Number of Exons, Relative Exon Number, Fraction of translatable transcripts	
					Maximum relative indel location, Minumum distance to exon boundary		

Feature Categories	VEST-indel In-frame	Sift In-frame	Ddig-in In-frame	VEST-indel Frameshift	Sift Frameshift	Ddig-in Frameshift	CADD
Computational predictions of protein local structure	High temperature factor	Whether indel is in a disordered region (RONN score)	Minimum disorder score	High temperature factor		Minimum disorder score, Max Coil Probability	
	Low stability			PredStabilityL			
	Solvent accessibility		Average, Maximum Accessible Surface Area	Solvent accessibility		Minimum Accessible Surface Area	
Protein local regional sequence composition	Positive formal charge						GC content
	Glutamine enriched						CpG
	Proline depletion			Proline depletion			Mapability
	Cysteine enriched						Motif based features
	Glycine enriched			Glycine enriched			Transcription Factor Binding Features
	Hydrophobic enrichment			Hydrophobic enrichment			Encode Features
	Aromatic enrichment						
	Negative formal charge						
	Shannon entropy of amino acids surrounding position	Whether indel resides in a repeat		Shannon entropy of amino acids surrounding position			
Amino acid residue conservation		Indel induced change to the HMM alignment score; Match to mach transition probability; Maximum HHBlits	Indel induced change to the HMM alignment score		Fraction of affected conserved amino acids	Position-based Protein Conservation score, Minimum PSSM conservation, Average HHBlits	Amino acid position from coding start, Relative position in coding sequence, Distance to splice, Reference Amino Acid, Sift/Polyphen scores
Protein-level annotations from UniProt				compositionally biased region			
	site of localization signal (protein targeted to secretory pathway or periplasm)						

Feature Categories	VEST-indel In-frame	Sift In-frame	Ddig-in In-frame	VEST-indel Frameshift	Sift Frameshift	Ddig-in Frameshift	CADD
	repeat region	Whether indel resides in a repeat					
	transmembrane domain						
					UniprotDOM_PPI		
Protein-domain-level annotations		Fraction of Pfam domains affected					

Candidate features are divided into eight categories: DNA sequence conservation, DNA natural variation, gene-level annotations, transcript-level annotations, computational predictions of protein local structure, protein local regional sequence composition, amino acid residue conservation, amino acid conservation, and protein-level annotations from UniProtKB. Only features used in the method's final classifier are included in the table. PROVEAN is not included in the table because it is a conservation based score and does not use features to make a classification.

Table S 4: List of all Possible Quaternary Boolean Combinations

1	X_1
2	X_2
3	X_3
4	X_4
5	$\left(X_1 \text{ AND } X_2\right)$
6	$\left(X_1 \text{ OR } X_2\right)$
7	$\left(X_1 \text{ AND } X_3\right)$
8	$\left(X_1 \text{ OR } X_3\right)$
9	$\left(X_1 \text{ AND } X_4\right)$
10	$\left(X_1 \text{ OR } X_4\right)$
11	$\left(X_2 \text{ AND } X_3\right)$
12	$\left(X_2 \text{ OR } X_3\right)$
13	$\left(X_2 \text{ AND } X_4\right)$
14	$\left(X_2 \text{ OR } X_4\right)$
15	$\left(X_3 \text{ AND } X_4\right)$
16	$\left(X_3 \text{ OR } X_4\right)$
17	$\left(X_1 \text{ AND } X_2 \text{ AND } X_3\right)$
18	$\left(X_1 \text{ OR } X_2 \text{ OR } X_3\right)$

19	$\left(X_1 \text{ AND } X_2 \text{ AND } X_4\right)$
20	$\left(X_1 \text{ OR } X_2 \text{ OR } X_4\right)$
21	$\left(X_1 \text{ AND } X_3 \text{ AND } X_4\right)$
22	$\left(X_1 \text{ OR } X_3 \text{ OR } X_4\right)$
23	$\left(X_2 \text{ AND } X_3 \text{ AND } X_4\right)$
24	$\left(X_2 \text{ OR } X_3 \text{ OR } X_4\right)$
25	$\left(\left(X_1 \text{ AND } X_2\right) \text{ OR } X_3\right)$
26	$\left(\left(X_1 \text{ OR } X_2\right) \text{ AND } X_3\right)$
27	$\left(\left(X_1 \text{ AND } X_3\right) \text{ OR } X_2\right)$
28	$\left(\left(X_1 \text{ OR } X_3\right) \text{ AND } X_2\right)$
29	$\left(\left(X_1 \text{ AND } X_2\right) \text{ OR } X_4\right)$
30	$\left(\left(X_1 \text{ OR } X_2\right) \text{ AND } X_4\right)$
31	$\left(\left(X_1 \text{ AND } X_4\right) \text{ OR } X_2\right)$
32	$\left(\left(X_1 \text{ OR } X_4\right) \text{ AND } X_2\right)$
33	$\left(\left(X_1 \text{ AND } X_3\right) \text{ OR } X_4\right)$
34	$\left(\left(X_1 \text{ OR } X_3\right) \text{ AND } X_4\right)$
35	$\left(\left(X_1 \text{ AND } X_4\right) \text{ OR } X_3\right)$
36	$\left(\left(X_1 \text{ OR } X_4\right) \text{ AND } X_3\right)$

37	$\left(\left(X_2 \text{ AND } X_3\right) \text{ OR } X_1\right)$
38	$\left(\left(X_2 \text{ OR } X_3\right) \text{ AND } X_1\right)$
39	$\left(\left(X_2 \text{ AND } X_4\right) \text{ OR } X_1\right)$
40	$\left(\left(X_2 \text{ OR } X_4\right) \text{ AND } X_1\right)$
41	$\left(\left(X_3 \text{ AND } X_4\right) \text{ OR } X_1\right)$
42	$\left(\left(X_3 \text{ OR } X_4\right) \text{ AND } X_1\right)$
43	$\left(\left(X_2 \text{ AND } X_3\right) \text{ OR } X_4\right)$
44	$\left(\left(X_2 \text{ OR } X_3\right) \text{ AND } X_4\right)$
45	$\left(\left(X_2 \text{ AND } X_4\right) \text{ OR } X_3\right)$
46	$\left(\left(X_2 \text{ OR } X_4\right) \text{ AND } X_3\right)$
47	$\left(\left(X_3 \text{ AND } X_4\right) \text{ OR } X_2\right)$
48	$\left(\left(X_3 \text{ OR } X_4\right) \text{ AND } X_4\right)$
49	$\left(X_1 \text{ AND } X_2 \text{ AND } X_3 \text{ AND } X_4\right)$
50	$\left(X_1 \text{ OR } X_2 \text{ OR } X_3 \square \text{ OR } X_4\right)$
51	$\left(\left(X_1 \text{ AND } X_2\right) \text{ OR } \left(X_3 \text{ AND } X_4\right)\right)$
52	$\left(\left(X_1 \text{ OR } X_2\right) \text{ AND } \left(X_3 \text{ OR } X_4\right)\right)$
53	$\left(\left(X_1 \text{ AND } X_3\right) \text{ OR } \left(X_2 \text{ AND } X_4\right)\right)$

54	$\left(\left(X_1 \text{ OR } X_3\right) \text{ AND } \left(X_2 \text{ OR } X_4\right)\right)$
55	$\left(\left(X_1 \text{ AND } X_4\right) \text{ OR } \left(X_2 \text{ AND } X_3\right)\right)$
56	$\left(\left(X_1 \text{ OR } X_4\right) \text{ AND } \left(X_2 \text{ OR } X_3\right)\right)$
57	$\left(X_1 \text{ OR } \left(X_2 \text{ AND } X_3 \text{ AND } X_4\right)\right)$
58	$\left(X_1 \text{ AND } \left(X_2 \text{ OR } X_3 \text{ OR } X_4\right)\right)$
59	$\left(X_2 \text{ OR } \left(X_1 \text{ AND } X_3 \text{ AND } X_4\right)\right)$
60	$\left(X_2 \text{ AND } \left(X_1 \text{ OR } X_3 \text{ OR } X_4\right)\right)$
61	$\left(X_3 \text{ OR } \left(X_1 \text{ AND } X_2 \text{ AND } X_4\right)\right)$
62	$\left(X_3 \text{ AND } \left(X_1 \text{ OR } X_2 \text{ OR } X_4\right)\right)$
63	$\left(X_4 \text{ OR } \left(X_1 \text{ AND } X_2 \text{ AND } X_3\right)\right)$
64	$\left(X_4 \text{ AND } \left(X_1 \text{ OR } X_2 \text{ OR } X_3\right)\right)$

Table S 5: Performance Comparison Restricted to Genes Containing at least one Pathogenic and one Benign Variant

In-frame	**Sensitivity**	**Specificity**	**Balanced Accuracy**
VEST-indel	0.743	0.815	0.779
SIFT-indel	0.680	0.770	0.725
PROVEAN	0.786	0.630	0.708
DDIG-in	0.627	0.870	0.749
CADD	0.745	0.727	0.736
Frameshift			
VEST-indel	0.536	0.796	0.666
SIFT-indel	0.934	0.174	0.554
DDIG-in	0.919	0.272	0.596
CADD	0.982	0.013	0.498

I compared the five tested methods on a difficult set of variants, limited to genes that contained at least one pathogenic and one benign variant (141 in-frame pathogenic and 78 benign variants in 57 genes and 561 frameshift pathogenic and 88 benign variants in 86 genes).

Table S 6: In-frame Boolean Meta-prediction Results

Method	Sensitivity	Specificity	Balanced Accuracy
(VEST-indel AND PROVEAN) OR (CADD AND DDIG-in)	0.930	0.974	0.952
(VEST-indel OR CADD) AND PROVEAN	0.947	0.955	0.951
(VEST-indel OR CADD) AND (PROVEAN OR DDIG-in)	0.947	0.949	0.948
VEST-indel OR (CADD AND PROVEAN AND DDIG-in	0.930	0.955	0.942
VEST-indel OR (CADD AND DDIG-in)	0.930	0.949	0.939
VEST-indel OR (DDIG-in AND CADD)	0.930	0.949	0.939
VEST-indel OR (CADD AND PROVEAN)	0.947	0.929	0.938
(VEST-indel OR DDIG-in) AND PROVEAN	0.930	0.942	0.936
(VEST-indel AND PROVEAN) OR (SIFT-indel AND DDIG-in)	0.912	0.955	0.934
(VEST-indel OR DDIG-in) AND (PROVEAN OR CADD)	0.930	0.936	0.933
(VEST-indel OR CADD OR DDIG-in) AND PROVEAN	0.947	0.917	0.932
(VEST-indel OR CADD) AND (PROVEAN OR SIFT-indel)	0.965	0.897	0.931
VEST-indel OR (SIFT-indel AND CADD AND DDIG-in	0.912	0.949	0.930
(VEST-indel OR SIFT-indel OR CADD) AND PROVEAN	0.947	0.910	0.929
VEST-indel OR (SIFT-indel AND PROVEAN AND CADD	0.912	0.942	0.927
VEST-indel OR (SIFT-indel AND PROVEAN AND DDIG-in	0.912	0.942	0.927
(VEST-indel AND PROVEAN) OR (SIFT-indel AND CADD)	0.930	0.923	0.926
VEST-indel OR (DDIG-in AND PROVEAN)	0.930	0.917	0.923
(VEST-indel OR DDIG-in) AND (PROVEAN OR SIFT-indel)	0.930	0.917	0.923
(VEST-indel AND DDIG-in) OR (PROVEAN AND CADD)	0.877	0.968	0.923
(VEST-indel AND SIFT-indel) OR	0.895	0.949	0.922

Method	Sensitivity	Specificity	Balanced Accuracy
(PROVEAN AND CADD)			
VEST-indel OR (SIFT-indel AND DDIG-in)	0.912	0.929	0.921
(VEST-indel OR SIFT-indel) AND PROVEAN	0.912	0.923	0.918
(VEST-indel AND PROVEAN) OR CADD	0.965	0.865	0.915
VEST-indel OR (SIFT-indel AND CADD)	0.930	0.897	0.914
(VEST-indel OR SIFT-indel OR DDIG-in) AND PROVEAN	0.930	0.897	0.914
(VEST-indel OR DDIG-in) AND (CADD OR SIFT-indel)	0.877	0.949	0.913
(VEST-indel OR SIFT-indel) AND (PROVEAN OR DDIG-in)	0.912	0.910	0.911
(VEST-indel AND PROVEAN) OR DDIG-in	0.930	0.891	0.910
(CADD AND PROVEAN) OR (SIFT-indel AND DDIG-in)	0.877	0.942	0.910
(VEST-indel AND CADD) OR (PROVEAN AND DDIG-in)	0.860	0.955	0.907
(CADD OR SIFT-indel) AND PROVEAN	0.895	0.917	0.906
(SIFT-indel OR CADD) AND PROVEAN	0.895	0.917	0.906
(VEST-indel AND CADD) OR (SIFT-indel AND DDIG-in)	0.842	0.968	0.905
VEST-indel OR (SIFT-indel AND PROVEAN)	0.912	0.897	0.905
(VEST-indel OR SIFT-indel) AND (PROVEAN OR CADD)	0.930	0.878	0.904
(CADD OR DDIG-in) AND PROVEAN	0.877	0.929	0.903
(VEST-indel OR PROVEAN) AND (CADD OR DDIG-in)	0.877	0.929	0.903
VEST-indel OR CADD	0.965	0.840	0.902
(VEST-indel AND SIFT-indel) OR (CADD AND DDIG-in)	0.825	0.974	0.899
(CADD OR SIFT-indel) AND (PROVEAN OR DDIG-in)	0.895	0.904	0.899
(VEST-indel OR PROVEAN) AND (SIFT-indel OR CADD)	0.895	0.904	0.899
VEST-indel OR DDIG-in	0.930	0.865	0.898

Method	Sensitivity	Specificity	Balanced Accuracy
(VEST-indel AND CADD) OR (PROVEAN AND SIFT-indel)	0.860	0.929	0.895
VEST-indel AND (CADD OR PROVEAN)	0.807	0.981	0.894
VEST-indel AND (CADD OR PROVEAN OR DDIG-in	0.807	0.981	0.894
VEST-indel AND (DDIG-in OR PROVEAN)	0.807	0.981	0.894
VEST-indel AND PROVEAN	0.807	0.981	0.894
(CADD OR SIFT-indel OR DDIG-in) AND PROVEAN	0.895	0.891	0.893
(VEST-indel AND SIFT-indel AND PROVEAN) OR CADD	0.912	0.872	0.892
(VEST-indel OR SIFT-indel) AND (CADD OR DDIG-in)	0.860	0.917	0.888
VEST-indel AND (SIFT-indel OR PROVEAN)	0.807	0.968	0.887
VEST-indel AND (SIFT-indel OR PROVEAN OR CADD	0.807	0.968	0.887
VEST-indel AND (SIFT-indel OR PROVEAN OR DDIG-in	0.807	0.968	0.887
(CADD OR DDIG-in) AND (PROVEAN OR SIFT-indel)	0.895	0.878	0.886
CADD OR (VEST-indel AND PROVEAN AND DDIG-in)	0.895	0.878	0.886
(VEST-indel AND DDIG-in) OR CADD	0.895	0.878	0.886
(VEST-indel AND SIFT-indel AND DDIG-in) OR CADD	0.895	0.878	0.886
(VEST-indel AND SIFT-indel) OR CADD	0.912	0.859	0.886
(VEST-indel AND SIFT-indel) OR (PROVEAN AND DDIG-in)	0.825	0.942	0.883
(VEST-indel OR CADD) AND DDIG-in	0.772	0.994	0.883
VEST-indel OR (CADD AND DDIG-in)	0.772	0.994	0.883
(VEST-indel OR CADD) AND (SIFT-indel OR DDIG-in)	0.842	0.923	0.883
(VEST-indel AND CADD AND PROVEAN) OR DDIG-in	0.860	0.904	0.882
(VEST-indel AND CADD) OR DDIG-in	0.860	0.904	0.882

Method	Sensitivity	Specificity	Balanced Accuracy
VEST-indel	0.807	0.955	0.881
CADD OR (SIFT-indel AND PROVEAN AND DDIG-in	0.895	0.865	0.880
(CADD AND PROVEAN) OR DDIG-in	0.877	0.878	0.878
DDIG-in OR (CADD AND PROVEAN)	0.877	0.878	0.878
(CADD AND DDIG-in) OR (PROVEAN AND SIFT-indel)	0.825	0.929	0.877
CADD OR (SIFT-indel AND DDIG-in)	0.895	0.859	0.877
(SIFT-indel AND DDIG-in) OR CADD	0.895	0.859	0.877
(VEST-indel AND DDIG-in) OR (CADD AND SIFT-indel)	0.807	0.942	0.875
PROVEAN	0.947	0.801	0.874
(VEST-indel AND CADD AND DDIG-in) OR PROVEAN	0.947	0.801	0.874
(VEST-indel AND CADD) OR PROVEAN	0.947	0.801	0.874
(VEST-indel AND DDIG-in) OR PROVEAN	0.947	0.801	0.874
(VEST-indel AND SIFT-indel AND CADD) OR PROVEAN	0.947	0.801	0.874
(VEST-indel AND SIFT-indel AND DDIG-in) OR PROVEAN	0.947	0.801	0.874
CADD OR (SIFT-indel AND DDIG-in)	0.772	0.974	0.873
DDIG-in AND (SIFT-indel OR CADD)	0.772	0.974	0.873
(VEST-indel OR SIFT-indel OR CADD) AND DDIG-in	0.772	0.974	0.873
(VEST-indel AND DDIG-in) OR (PROVEAN AND SIFT-indel)	0.807	0.936	0.871
(CADD AND DDIG-in) OR PROVEAN	0.947	0.795	0.871
(CADD AND SIFT-indel AND DDIG-in) OR PROVEAN	0.947	0.795	0.871
CADD OR (SIFT-indel AND PROVEAN)	0.912	0.827	0.870
(SIFT-indel AND DDIG-in) OR PROVEAN	0.947	0.788	0.868
(VEST-indel AND SIFT-indel) OR	0.947	0.788	0.868

Method	Sensitivity	Specificity	Balanced Accuracy
PROVEAN			
(SIFT-indel OR DDIG-in) AND PROVEAN	0.825	0.910	0.867
CADD OR (DDIG-in AND PROVEAN)	0.895	0.840	0.867
(DDIG-in AND PROVEAN) OR CADD	0.895	0.840	0.867
DDIG-in AND PROVEAN	0.772	0.962	0.867
(VEST-indel OR PROVEAN) AND DDIG-in	0.772	0.962	0.867
VEST-indel AND (CADD OR DDIG-in)	0.737	0.994	0.865
VEST-indel AND (DDIG-in OR CADD)	0.737	0.994	0.865
VEST-indel AND (SIFT-indel OR CADD)	0.754	0.974	0.864
VEST-indel AND (SIFT-indel OR CADD OR DDIG-in	0.754	0.974	0.864
VEST-indel OR (SIFT-indel AND DDIG-in)	0.754	0.974	0.864
(CADD AND SIFT-indel) OR (PROVEAN AND DDIG-in)	0.825	0.904	0.864
(VEST-indel AND SIFT-indel AND PROVEAN) OR DDIG-in	0.825	0.904	0.864
(CADD OR PROVEAN) AND DDIG-in	0.772	0.955	0.864
DDIG-in AND (CADD OR PROVEAN)	0.772	0.955	0.864
(VEST-indel OR CADD OR PROVEAN) AND DDIG-in	0.772	0.955	0.864
(VEST-indel OR DDIG-in) AND SIFT-indel	0.772	0.955	0.864
VEST-indel OR PROVEAN	0.947	0.776	0.862
(VEST-indel OR PROVEAN) AND (SIFT-indel OR DDIG-in)	0.825	0.897	0.861
(CADD AND SIFT-indel) OR PROVEAN	0.965	0.756	0.861
(SIFT-indel AND CADD) OR PROVEAN	0.965	0.756	0.861
VEST-indel OR CADD OR DDIG-in	0.965	0.756	0.861
VEST-indel OR DDIG-in OR CADD	0.965	0.756	0.861

Method	Sensitivity	Specificity	Balanced Accuracy
(CADD OR SIFT-indel OR PROVEAN) AND DDIG-in	0.772	0.949	0.860
DDIG-in AND (SIFT-indel OR PROVEAN)	0.772	0.949	0.860
(VEST-indel OR SIFT-indel OR PROVEAN) AND DDIG-in	0.772	0.949	0.860
(VEST-indel AND SIFT-indel AND CADD) OR DDIG-in	0.807	0.910	0.859
(VEST-indel AND SIFT-indel) OR DDIG-in	0.825	0.891	0.858
(VEST-indel OR CADD) AND SIFT-indel	0.789	0.923	0.856
(VEST-indel AND PROVEAN) OR SIFT-indel	0.965	0.744	0.854
SIFT-indel AND PROVEAN	0.772	0.936	0.854
(CADD OR PROVEAN) AND (SIFT-indel OR DDIG-in)	0.842	0.865	0.854
SIFT-indel AND PROVEAN AND DDIG-in	0.719	0.987	0.853
(CADD AND SIFT-indel AND PROVEAN) OR DDIG-in	0.807	0.897	0.852
DDIG-in OR PROVEAN	0.947	0.750	0.849
VEST-indel OR SIFT-indel	0.965	0.731	0.848
(CADD OR DDIG-in) AND SIFT-indel	0.772	0.923	0.848
(DDIG-in OR CADD) AND SIFT-indel	0.772	0.923	0.848
(DDIG-in OR PROVEAN) AND SIFT-indel	0.772	0.923	0.848
(VEST-indel OR PROVEAN) AND SIFT-indel	0.772	0.923	0.848
SIFT-indel AND DDIG-in	0.719	0.974	0.847
SIFT-indel AND (VEST-indel OR CADD OR DDIG-in)	0.789	0.904	0.847
CADD OR DDIG-in	0.895	0.795	0.845
DDIG-in OR CADD	0.895	0.795	0.845
(VEST-indel OR DDIG-in) AND CADD	0.702	0.987	0.844
CADD AND PROVEAN	0.719	0.968	0.844
(VEST-indel OR PROVEAN) AND CADD	0.719	0.968	0.844
(CADD AND PROVEAN) OR	0.947	0.737	0.842

Method	Sensitivity	Specificity	Balanced Accuracy
SIFT-indel			
(CADD AND SIFT-indel) OR DDIG-in	0.825	0.859	0.842
DDIG-in OR (SIFT-indel AND CADD)	0.825	0.859	0.842
DDIG-in OR (SIFT-indel AND PROVEAN)	0.825	0.859	0.842
VEST-indel AND (SIFT-indel OR DDIG-in)	0.702	0.981	0.841
DDIG-in	0.772	0.910	0.841
SIFT-indel AND (VEST-indel OR PROVEAN OR DDIG-in)	0.772	0.910	0.841
CADD AND (DDIG-in OR PROVEAN)	0.719	0.962	0.840
CADD AND (VEST-indel OR PROVEAN OR DDIG-in)	0.719	0.962	0.840
(DDIG-in OR PROVEAN) AND CADD	0.719	0.962	0.840
(CADD OR PROVEAN) AND SIFT-indel	0.789	0.891	0.840
CADD OR PROVEAN	0.965	0.712	0.838
SIFT-indel AND (CADD OR PROVEAN OR DDIG-in)	0.789	0.885	0.837
VEST-indel OR SIFT-indel OR CADD	1.000	0.673	0.837
VEST-indel OR DDIG-in OR PROVEAN	0.947	0.724	0.836
SIFT-indel AND (VEST-indel OR PROVEAN OR CADD)	0.789	0.878	0.834
SIFT-indel OR (VEST-indel AND PROVEAN AND CADD)	0.912	0.750	0.831
(VEST-indel AND CADD) OR SIFT-indel	0.912	0.750	0.831
VEST-indel AND SIFT-indel AND PROVEAN	0.667	0.994	0.830
CADD AND (SIFT-indel OR PROVEAN)	0.737	0.923	0.830
CADD AND (SIFT-indel OR PROVEAN OR DDIG-in	0.737	0.923	0.830
(VEST-indel OR SIFT-indel OR PROVEAN) AND CADD	0.737	0.923	0.830
(VEST-indel OR SIFT-indel OR DDIG-in) AND CADD	0.719	0.936	0.828

Method	Sensitivity	Specificity	Balanced Accuracy
VEST-indel OR CADD OR PROVEAN	0.965	0.686	0.825
VEST-indel AND DDIG-in	0.649	1.000	0.825
VEST-indel AND DDIG-in AND PROVEAN	0.649	1.000	0.825
VEST-indel OR SIFT-indel OR DDIG-in	0.982	0.667	0.825
VEST-indel AND SIFT-indel	0.667	0.981	0.824
CADD OR SIFT-indel	0.947	0.692	0.820
SIFT-indel OR CADD	0.947	0.692	0.820
(VEST-indel OR SIFT-indel) AND CADD	0.702	0.936	0.819
VEST-indel OR (SIFT-indel AND CADD)	0.702	0.936	0.819
(CADD AND DDIG-in) OR SIFT-indel	0.877	0.756	0.817
(DDIG-in AND CADD) OR SIFT-indel	0.877	0.756	0.817
SIFT-indel OR (CADD AND PROVEAN AND DDIG-in)	0.877	0.756	0.817
CADD OR DDIG-in OR PROVEAN	0.965	0.667	0.816
CADD OR PROVEAN OR DDIG-in	0.965	0.667	0.816
SIFT-indel OR PROVEAN	1.000	0.622	0.811
SIFT-indel OR (VEST-indel AND CADD AND DDIG-in)	0.860	0.756	0.808
SIFT-indel OR (VEST-indel AND PROVEAN AND DDIG-in)	0.860	0.756	0.808
(VEST-indel AND DDIG-in) OR SIFT-indel	0.860	0.756	0.808
CADD	0.737	0.878	0.808
CADD AND DDIG-in AND PROVEAN	0.614	1.000	0.807
CADD AND PROVEAN AND DDIG-in	0.614	1.000	0.807
VEST-indel AND SIFT-indel AND DDIG-in	0.614	1.000	0.807
VEST-indel AND SIFT-indel AND PROVEAN AND DDIG-in	0.614	1.000	0.807
VEST-indel OR SIFT-indel OR CADD OR DDIG-in	1.000	0.609	0.804

Method	Sensitivity	Specificity	Balanced Accuracy
VEST-indel OR SIFT-indel OR PROVEAN	1.000	0.609	0.804
CADD AND (SIFT-indel OR DDIG-in)	0.667	0.942	0.804
(SIFT-indel OR DDIG-in) AND CADD	0.667	0.942	0.804
(DDIG-in AND PROVEAN) OR SIFT-indel	0.877	0.731	0.804
CADD AND DDIG-in	0.614	0.994	0.804
DDIG-in AND CADD	0.614	0.994	0.804
VEST-indel OR CADD OR PROVEAN OR DDIG-in	0.965	0.641	0.803
CADD AND SIFT-indel AND PROVEAN	0.596	0.987	0.792
SIFT-indel AND PROVEAN AND CADD	0.596	0.987	0.792
SIFT-indel OR PROVEAN OR DDIG-in	1.000	0.583	0.792
SIFT-indel	0.825	0.756	0.790
CADD OR SIFT-indel OR PROVEAN	1.000	0.577	0.788
SIFT-indel OR PROVEAN OR CADD	1.000	0.577	0.788
CADD OR SIFT-indel OR DDIG-in	0.947	0.628	0.788
SIFT-indel OR CADD OR DDIG-in	0.947	0.628	0.788
VEST-indel AND CADD	0.579	0.994	0.786
VEST-indel AND CADD AND PROVEAN	0.579	0.994	0.786
VEST-indel OR SIFT-indel OR PROVEAN OR DDIG-in	1.000	0.571	0.785
SIFT-indel OR DDIG-in	0.877	0.692	0.785
VEST-indel OR SIFT-indel OR PROVEAN OR CADD	1.000	0.564	0.782
CADD AND SIFT-indel AND PROVEAN AND DDIG-in	0.561	1.000	0.781
CADD AND SIFT-indel	0.614	0.942	0.778
SIFT-indel AND CADD	0.614	0.942	0.778
CADD AND SIFT-indel AND DDIG-in	0.561	0.994	0.777
SIFT-indel AND CADD AND DDIG-in	0.561	0.994	0.777
CADD OR SIFT-indel OR	1.000	0.538	0.769

Method	Sensitivity	Specificity	Balanced Accuracy
PROVEAN OR DDIG-in			
VEST-indel AND CADD AND DDIG-in	0.491	1.000	0.746
VEST-indel AND CADD AND PROVEAN AND DDIG-in	0.491	1.000	0.746
VEST-indel AND DDIG-in AND CADD	0.491	1.000	0.746
VEST-indel AND SIFT-indel AND CADD	0.491	1.000	0.746
VEST-indel AND SIFT-indel AND PROVEAN AND CADD	0.491	1.000	0.746
VEST-indel AND SIFT-indel AND CADD AND DDIG-in	0.456	1.000	0.728

Table S 7: ***Frameshift Boolean Meta-prediction Results***

Method	Sensitivity	Specificity	Balanced Accuracy
VEST-indel AND (SIFT-indel OR DDIG-in)	0.835	0.967	0.901
VEST-indel	0.849	0.950	0.900
VEST-indel AND (SIFT-indel OR CADD)	0.849	0.950	0.900
VEST-indel AND (SIFT-indel OR CADD OR DDIG-in	0.849	0.950	0.900
VEST-indel AND (DDIG-in OR CADD)	0.841	0.950	0.896
VEST-indel AND CADD	0.833	0.950	0.891
VEST-indel AND SIFT-indel	0.797	0.967	0.882
VEST-indel AND SIFT-indel AND CADD	0.780	0.967	0.874
VEST-indel OR (SIFT-indel AND DDIG-in)	0.946	0.783	0.864
VEST-indel OR (SIFT-indel AND CADD AND DDIG-in	0.946	0.783	0.864
(VEST-indel AND SIFT-indel) OR DDIG-in	0.941	0.783	0.862
(VEST-indel AND SIFT-indel) OR (CADD AND DDIG-in)	0.941	0.783	0.862
VEST-indel OR DDIG-in	0.956	0.767	0.861
VEST-indel OR (DDIG-in AND CADD)	0.956	0.767	0.861
(VEST-indel OR DDIG-in) AND (CADD OR SIFT-indel)	0.956	0.767	0.861
(VEST-indel AND CADD) OR (SIFT-indel AND DDIG-in)	0.937	0.783	0.860
(VEST-indel AND SIFT-indel AND CADD) OR DDIG-in	0.933	0.783	0.858
(VEST-indel AND CADD) OR DDIG-in	0.948	0.767	0.857
(VEST-indel OR DDIG-in) AND CADD	0.939	0.767	0.853
(VEST-indel OR DDIG-in) AND SIFT-indel	0.893	0.800	0.847
VEST-indel AND DDIG-in	0.638	0.983	0.811
VEST-indel AND DDIG-in AND CADD	0.630	0.983	0.807
VEST-indel AND SIFT-indel AND DDIG-in	0.600	0.983	0.792
VEST-indel AND SIFT-indel AND	0.592	0.983	0.788

Method	Sensitivity	Specificity	Balanced Accuracy
CADD AND DDIG-in			
VEST-indel OR (SIFT-indel AND DDIG-in)	0.734	0.817	0.775
DDIG-in	0.745	0.800	0.772
(VEST-indel OR CADD) AND DDIG-in	0.745	0.800	0.772
DDIG-in AND (SIFT-indel OR CADD)	0.745	0.800	0.772
(VEST-indel OR SIFT-indel OR CADD) AND DDIG-in	0.745	0.800	0.772
DDIG-in AND CADD	0.736	0.800	0.768
SIFT-indel AND DDIG-in	0.697	0.817	0.757
SIFT-indel AND CADD AND DDIG-in	0.688	0.817	0.752
VEST-indel OR (SIFT-indel AND CADD)	0.987	0.283	0.635
(VEST-indel OR CADD) AND (SIFT-indel OR DDIG-in)	0.983	0.283	0.633
(VEST-indel AND DDIG-in) OR (CADD AND SIFT-indel)	0.964	0.300	0.632
(VEST-indel OR SIFT-indel) AND (CADD OR DDIG-in)	0.979	0.283	0.631
DDIG-in OR (SIFT-indel AND CADD)	0.975	0.283	0.629
(VEST-indel OR SIFT-indel) AND CADD	0.971	0.283	0.627
(SIFT-indel OR DDIG-in) AND CADD	0.967	0.283	0.625
(VEST-indel OR SIFT-indel OR DDIG-in) AND CADD	0.981	0.267	0.624
(VEST-indel OR CADD) AND SIFT-indel	0.935	0.300	0.618
SIFT-indel AND (VEST-indel OR CADD OR DDIG-in)	0.935	0.300	0.618
(DDIG-in OR CADD) AND SIFT-indel	0.927	0.300	0.613
(VEST-indel AND DDIG-in) OR SIFT-indel	0.973	0.250	0.611
SIFT-indel OR (VEST-indel AND CADD AND DDIG-in)	0.973	0.250	0.611
VEST-indel OR SIFT-indel	0.987	0.233	0.610
(VEST-indel AND CADD) OR SIFT-indel	0.987	0.233	0.610

Method	Sensitivity	Specificity	Balanced Accuracy
SIFT-indel AND CADD	0.918	0.300	0.609
SIFT-indel OR DDIG-in	0.983	0.233	0.608
(DDIG-in AND CADD) OR SIFT-indel	0.983	0.233	0.608
VEST-indel OR SIFT-indel OR DDIG-in	0.998	0.217	0.607
SIFT-indel	0.935	0.250	0.593
VEST-indel OR CADD	1.000	0.050	0.525
(VEST-indel AND SIFT-indel) OR CADD	1.000	0.050	0.525
VEST-indel OR DDIG-in OR CADD	1.000	0.050	0.525
DDIG-in OR CADD	0.992	0.050	0.521
(VEST-indel AND DDIG-in) OR CADD	0.992	0.050	0.521
(SIFT-indel AND DDIG-in) OR CADD	0.992	0.050	0.521
(VEST-indel AND SIFT-indel AND DDIG-in) OR CADD	0.992	0.050	0.521
CADD	0.983	0.050	0.517
SIFT-indel OR CADD	1.000	- 0.000	0.500
VEST-indel OR SIFT-indel OR CADD	1.000	- 0.000	0.500
SIFT-indel OR CADD OR DDIG-in	1.000	- 0.000	0.500
VEST-indel OR SIFT-indel OR CADD OR DDIG-in	1.000	- 0.000	0.500

Bibliography

1. Kumar, P., Henikoff, S., and Ng, P.C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nature protocols 4, 1073-1081.
2. Reva, B., Antipin, Y., and Sander, C. (2011). Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic acids research, gkr407.
3. Stone, E.A., and Sidow, A. (2005). Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. Genome research 15, 978-986.
4. Thomas, P.D., Campbell, M.J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., Diemer, K., Muruganujan, A., and Narechania, A. (2003). PANTHER: a library of protein families and subfamilies indexed by function. Genome research 13, 2129-2141.
5. Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. Nature methods 7, 248-249.
6. Bromberg, Y., and Rost, B. (2007). SNAP: predict effect of non-synonymous polymorphisms on function. Nucleic acids research 35, 3823-3835.
7. Li, B., Krishnan, V.G., Mort, M.E., Xin, F., Kamati, K.K., Cooper, D.N., Mooney, S.D., and Radivojac, P. (2009). Automated inference of molecular mechanisms of disease from amino acid substitutions. Bioinformatics 25, 2744-2750.
8. Schwarz, J.M., Rödelsperger, C., Schuelke, M., and Seelow, D. (2010). MutationTaster evaluates disease-causing potential of sequence alterations. Nature methods 7, 575-576.
9. Ng, P.C., and Henikoff, S. (2006). Predicting the effects of amino acid substitutions on protein function. Annu Rev Genomics Hum Genet 7, 61-80.
10. Cooper, G.M., and Shendure, J. (2011). Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. Nature Reviews Genetics 12, 628-640.
11. Lopez-Bigas, N., Audit, B., Ouzounis, C., Parra, G., and Guig√≥, R. (2005). Are splicing mutations the most frequent cause of hereditary disease? FEBS Letters 579, 1900-1903.
12. Krawczak, M., Thomas, N.S., Hundrieser, B., Mort, M., Wittig, M., Hampe, J., and Cooper, D.N. (2006). Single base-pair substitutions in exon-intron junctions of human genes: nature, distribution, and consequences for mRNA splicing. Human Mutation 28, 150-158.
13. Sterne-Weiler, T., Howard, J., Mort, M., Cooper, D.N., and Sanford, J.R. (2011). Loss of exon identity is a common mechanism of human inherited disease. Genome research 21, 1563-1571.
14. Lim, K.H., Ferraris, L., Filloux, M.E., Raphael, B.J., and Fairbrother, W.G. (2011). Using positional distribution to identify splicing elements and predict pre-mRNA

processing defects in human genes. Proceedings of the National Academy of Sciences 108, 11093-11098.
15. Pagani, F., and Baralle, F.E. (2004). Genomic variants in exons and introns: identifying the splicing spoilers. Nature Reviews Genetics 5, 389-396.
16. Shapiro, M.B., and Senapathy, P. (1987). RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression. Nucleic acids research 15, 7155-7174.
17. Desmet, F.ß.-O., Hamroun, D., Lalande, M., Collod-B√©roud, G.l., Claustres, M., and B√©roud, C. (2009). Human Splicing Finder: an online bioinformatics tool to predict splicing signals. Nucleic Acids Research 37, e67-e67.
18. Yeo, G., and Burge, C.B. (2004). Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. Journal of Computational Biology 11, 377-394.
19. Reese, M.G., Eeckman, F.H., Kulp, D., and Haussler, D. (1997). Improved splice site detection in Genie. Journal of Computational Biology 4, 311-323.
20. Keren, H., Lev-Maor, G., and Ast, G. (2010). Alternative splicing and evolution: diversification, exon definition and function. Nature Reviews Genetics 11, 345-355.
21. Lim, K.H., and Fairbrother, W.G. (2012). Spliceman,Äîa computational web server that predicts sequence variations in pre-mRNA splicing. Bioinformatics 28, 1031-1032.
22. Fairbrother, W.G., Yeo, G.W., Yeh, R., Goldstein, P., Mawson, M., Sharp, P.A., and Burge, C.B. (2004). RESCUE-ESE identifies candidate exonic splicing enhancers in vertebrate exons. Nucleic Acids Research 32, W187-W190.
23. Hiller, M., Zhang, Z., Backofen, R., and Stamm, S. (2007). Pre-mRNA secondary structures influence exon recognition. PLoS Genetics 3, e204.
24. Consortium, G.P. (2010). A map of human genome variation from population-scale sequencing. Nature 467, 1061-1073.
25. Steward, R.E., MacArthur, M.W., Laskowski, R.A., and Thornton, J.M. (2003). Molecular basis of inherited diseases: a structural perspective. Trends in Genetics 19, 505-513.
26. Ribic, C.M., Sargent, D.J., Moore, M.J., Thibodeau, S.N., French, A.J., Goldberg, R.M., Hamilton, S.R., Laurent-Puig, P., Gryfe, R., and Shepherd, L.E. (2003). Tumor microsatellite-instability status as a predictor of benefit from fluorouracil-based adjuvant chemotherapy for colon cancer. New England Journal of Medicine 349, 247-257.
27. Boland, C.R., Thibodeau, S.N., Hamilton, S.R., Sidransky, D., Eshleman, J.R., Burt, R.W., Meltzer, S.J., Rodriguez-Bigas, M.A., Fodde, R., and Ranzani, G.N. (1998). A National Cancer Institute Workshop on Microsatellite Instability for cancer detection and familial predisposition: development of international criteria for the determination of microsatellite instability in colorectal cancer. In. (AACR.
28. Treangen, T.J., and Salzberg, S.L. (2012). Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nature Reviews Genetics 13, 36-46.
29. Harper, P.S. (2010). Practical Genetic Counselling 7th Edition.(CRC Press).

30. Boveri, T. (2008). Concerning the origin of malignant tumours by Theodor Boveri. Translated and annotated by Henry Harris. Journal of cell science 121, 1-84.
31. Lengauer, C., Kinzler, K.W., and Vogelstein, B. (1998). Genetic instabilities in human cancers. Nature 396, 643-649.
32. Duesberg, P., and Li, R. (2003). Multistep carcinogenesis: a chain reaction of aneuploidizations. Cell cycle 2, 201-209.
33. Tabor, A., Madsen, M., Obel, E., Philip, J., Bang, J., and Gaard-Pedersen, B.r. (1986). Randomised controlled trial of genetic amniocentesis in 4606 low-risk women. The Lancet 327, 1287-1293.
34. Rietbergen, J.B., Kruger, A.E.B., Kranse, R., and Schröder, F.H. (1997). Complications of transrectal ultrasound-guided systematic sextant biopsies of the prostate: evaluation of complication rates and risk factors within a population-based screening program. Urology 49, 875-880.
35. Swaminathan, R., and Butt, A.N. (2006). Circulating nucleic acids in plasma and serum. Annals of the New York Academy of Sciences 1075, 1-9.
36. Lo, Y.D., Corbetta, N., Chamberlain, P.F., Rai, V., Sargent, I.L., Redman, C.W., and Wainscoat, J.S. (1997). Presence of fetal DNA in maternal plasma and serum. The Lancet 350, 485-487.
37. Schwarzenbach, H., Hoon, D.S., and Pantel, K. (2011). Cell-free nucleic acids as biomarkers in cancer patients. Nature Reviews Cancer 11, 426-437.
38. Norton, M.E., Brar, H., Weiss, J., Karimi, A., Laurent, L.C., Caughey, A.B., Rodriguez, M.H., Williams, J., Mitchell, M.E., and Adair, C.D. (2012). Non-Invasive Chromosomal Evaluation (NICE) Study: results of a multicenter prospective cohort study for detection of fetal trisomy 21 and trisomy 18. American journal of obstetrics and gynecology 207, 137. e131-137. e138.
39. Bianchi, D.W., Platt, L.D., Goldberg, J.D., Abuhamad, A.Z., Sehnert, A.J., and Rava, R.P. (2012). Genome-wide fetal aneuploidy detection by maternal plasma DNA sequencing. Obstetrics & Gynecology 119, 890-901.
40. Nicolaides, K.H., Syngelaki, A., Ashoor, G., Birdir, C., and Touzet, G. (2012). Noninvasive prenatal testing for fetal trisomies in a routinely screened first-trimester population. American journal of obstetrics and gynecology 207, 374. e371-374. e376.
41. Diehl, F., Schmidt, K., Choti, M.A., Romans, K., Goodman, S., Li, M., Thornton, K., Agrawal, N., Sokoll, L., and Szabo, S.A. (2008). Circulating mutant DNA to assess tumor dynamics. Nature medicine 14, 985-990.
42. Benjamini, Y., and Speed, T.P. (2012). Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic acids research, gks001.
43. Sathirapongsasuti, J.F., Lee, H., Horst, B.A., Brunner, G., Cochran, A.J., Binder, S., Quackenbush, J., and Nelson, S.F. (2011). Exome sequencing-based copy-number variation and loss of heterozygosity detection: ExomeCNV. Bioinformatics 27, 2648-2654.
44. Plagnol, V., Curtis, J., Epstein, M., Mok, K.Y., Stebbings, E., Grigoriadou, S., Wood, N.W., Hambleton, S., Burns, S.O., and Thrasher, A.J. (2012). A robust model for read count data in exome sequencing experiments and implications for copy number variant calling. Bioinformatics 28, 2747-2754.

45. Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., and FitzHugh, W. (2001). Initial sequencing and analysis of the human genome. Nature 409, 860-921.
46. Zhao, M., Wang, Q., Wang, Q., Jia, P., and Zhao, Z. (2013). Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. BMC bioinformatics 14, S1.
47. Wou, K., Feinberg, J.L., Wapner, R.J., and Simpson, J.L. (2015). Cell-free DNA versus intact fetal cells for prenatal genetic diagnostics: what does the future hold? Expert review of molecular diagnostics 15, 989-998.
48. Chan, L.L., and Jiang, P. (2015). Bioinformatics analysis of circulating cell-free DNA sequencing data. Clinical biochemistry 48, 962-975.
49. Choi, Y., Sims, G.E., Murphy, S., Miller, J.R., and Chan, A.P. (2012). Predicting the functional effect of amino acid substitutions and indels. PloS ONE 7, e46688.
50. Folkman, L., Yang, Y., Li, Z., Stantic, B., Sattar, A., Mort, M., Cooper, D.N., Liu, Y., and Zhou, Y. (2015). DDIG-in: detecting disease-causing genetic variations due to frameshifting indels and nonsense mutations employing sequence and structural properties at nucleotide and protein levels. Bioinformatics 31, p.1599-1606.
51. Hu, J., and Ng, P.C. (2012). Predicting the effects of frameshifting indels. Genome Biol 13, R9.
52. Hu, J., and Ng, P.C. (2013). SIFT Indel: predictions for the functional effects of amino acid insertions/deletions in proteins. PloS ONE 8, e77940.
53. Zhao, H., Yang, Y., Lin, H., Zhang, X., Mort, M., Cooper, D.N., Liu, Y., and Zhou, Y. (2013). DDIG-in: discriminating between disease-associated and neutral non-frameshifting micro-indels. Genome Biol 14, R23.
54. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I.H. (2009). The WEKA data mining software: an update. ACM SIGKDD explorations newsletter 11, 10-18.
55. Kircher, M., Witten, D.M., Jain, P., O'roak, B.J., Cooper, G.M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. Nature genetics 46, 310-315.
56. Stenson, P.D., Mort, M., Ball, E.V., Shaw, K., Phillips, A.D., and Cooper, D.N. (2014). The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. Human genetics 133, 1-9.
57. Fu, W., O'Connor, T.D., Jun, G., Kang, H.M., Abecasis, G., Leal, S.M., Gabriel, S., Rieder, M.J., Altshuler, D., and Shendure, J. (2013). Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature 493, 216-220.
58. Clarke, L., Zheng-Bradley, X., Smith, R., Kulesha, E., Xiao, C., Toneva, I., Vaughan, B., Preuss, D., Leinonen, R., and Shumway, M. (2012). The 1000 Genomes Project: data management and community access. Nature Methods 9, 459-462.
59. Lohmueller, K.E., Indap, A.R., Schmidt, S., Boyko, A.R., Hernandez, R.D., Hubisz, M.J., Sninsky, J.J., White, T.J., Sunyaev, S.R., and Nielsen, R. (2008). Proportionally more deleterious genetic variation in European than in African populations. Nature 451, 994-997.

60. MacArthur, D.G., and Tyler-Smith, C. (2010). Loss-of-function variants in the genomes of healthy humans. Human Molecular Genetics 19, R125-R130.
61. Ng, P.C., Levy, S., Huang, J., Stockwell, T.B., Walenz, B.P., Li, K., Axelrod, N., Busam, D.A., Strausberg, R.L., and Venter, J.C. (2008). Genetic variation in an individual human exome. PLoS Genetics 4, e1000160.
62. Breiman, L. (2001). Random forests. Machine Learning 45, 5-32.
63. Wong, W.C., Kim, D., Carter, H., Diekhans, M., Ryan, M.C., and Karchin, R. (2011). CHASM and SNVBox: toolkit for detecting biologically important single nucleotide mutations in cancer. Bioinformatics 27, 2147-2148.
64. Knijnenburg, T.A., Wessels, L.F., Reinders, M.J., and Shmulevich, I. (2009). Fewer permutations, more accurate P-values. Bioinformatics 25, i161-i168.
65. Pickands III, J. (1975). Statistical inference using extreme order statistics. the Annals of Statistics, 119-131.
66. Landrum, M.J., Lee, J.M., Riley, G.R., Jang, W., Rubinstein, W.S., Church, D.M., and Maglott, D.R. (2013). ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Research, gkt1113.
67. Altschul, S.F., Madden, T.L., Schaeffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research 25, 3389-3402.
68. Capriotti, E., and Altman, R.B. (2011). A new disease-specific machine learning approach for the prediction of cancer-causing missense variants. Genomics 98, 310.
69. Frousios, K., Iliopoulos, C.S., Schlitt, T., and Simpson, M.A. (2013). Predicting the functional consequences of non-synonymous DNA sequence variants‚ÄîÄîevaluation of bioinformatics tools and development of a consensus strategy. Genomics 102, 223-228.
70. Gonzalez-Perez, A., and Lopez-Bigas, N. (2011). Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. The American Journal of Human Genetics 88, 440-449.
71. Martelotto, L.G., Ng, C.K., De Filippo, M.R., Zhang, Y., Piscuoglio, S., Lim, R., Shen, R., Norton, L., Reis-Filho, J.S., and Weigelt, B. (2014). Benchmarking mutation effect prediction algorithms using functionally validated cancer-related missense mutations. Genome biology 15, 484.
72. Fawcett, T. (2004). ROC graphs: Notes and practical considerations for researchers. Machine Learning 31, 1-38.
73. Carter, H., Douville, C., Stenson, P.D., Cooper, D.N., and Karchin, R. (2013). Identifying Mendelian disease genes with the variant effect scoring tool. BMC Genomics 14, S3.
74. Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123, 585-595.
75. Pruitt, K.D., Tatusova, T., and Maglott, D.R. (2007). NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic acids research 35, D61-D65.
76. Cunningham, F., Amode, M.R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., and Fitzgerald, S. (2015). Ensembl 2015. Nucleic Acids Research 43, D662-D669.

77. Pfeifer, B., Wittelsburger, U., Onsins, S.E.R., and Lercher, M.J. (2014). PopGenome: an efficient Swiss army knife for population genomic analyses in R. Molecular Biology and Evolution, msu136.
78. Tennessen, J.A., Madeoy, J., and Akey, J.M. (2010). Signatures of positive selection apparent in a small sample of human exomes. Genome Research 20, 1327-1334.
79. Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M., and Bairoch, A. (2007). Uniprotkb/swiss-prot. In Plant Bioinformatics. (Springer), pp 89-112.
80. Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R., and Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. Genome research 20, 110-121.
81. Chan, P.A., Duraisamy, S., Miller, P.J., Newell, J.A., McBride, C., Bond, J.P., Raevaara, T., Ollila, S., Nystr√ ∂ m, M., and Grimm, A.J. (2007). Interpreting missense variants: comparing computational methods in human disease genes CDKN2A, MLH1, MSH2, MECP2, and tyrosinase (TYR). Human Mutation 28, 683-693.
82. Hicks, S., Wheeler, D.A., Plon, S.E., and Kimmel, M. (2011). Prediction of missense mutation functionality depends on both the algorithm and sequence alignment employed. Human Mutation 32, 661-668.
83. Shihab, H.A., Gough, J., Cooper, D.N., Stenson, P.D., Barker, G.L., Edwards, K.J., Day, I.N., and Gaunt, T.R. (2013). Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. Human Mutation 34, 57-65.
84. Thusberg, J., Olatubosun, A., and Vihinen, M. (2011). Performance of mutation pathogenicity prediction methods on missense variants. Human Mutation 32, 358-368.
85. Douville, C., Carter, H., Kim, R., Niknafs, N., Diekhans, M., Stenson, P.D., Cooper, D.N., Ryan, M., and Karchin, R. (2013). CRAVAT: cancer-related analysis of variants toolkit. Bioinformatics 29, 647-648.
86. Benn, P.A., and Hsu, L.Y.F. (2004). Prenatal diagnosis of chromosomal abnormalities through amniocentesis.(The John Hopkins University Press: Baltimore).
87. Staebler, M., Donner, C., Van Regemorter, N., Duprez, L., De Maertelaer, V., Devreker, F., and Avni, F. (2005). Should determination of the karyotype be systematic for all malformations detected by obstetrical ultrasound? Prenatal diagnosis 25, 567-573.
88. Wellesley, D., Dolk, H., Boyd, P.A., Greenlees, R., Haeusler, M., Nelen, V., Garne, E., Khoshnood, B., Doray, B., and Rissmann, A. (2012). Rare chromosome abnormalities, prevalence and prenatal diagnosis rates from population-based congenital anomaly registers in Europe. European Journal of Human Genetics 20, 521-526.
89. Obstetricians, A.C.o., and Gynecologists. (2007). ACOG Practice Bulletin No. 88, December 2007. Invasive prenatal testing for aneuploidy. Obstetrics and gynecology 110, 1459.
90. Gil, M., Quezada, M., Revello, R., Akolekar, R., and Nicolaides, K. (2015). Analysis of cell‐free DNA in maternal blood in screening for fetal aneuploidies: updated meta‐analysis. Ultrasound in obstetrics & gynecology 45, 249-266.

91. Bianchi, D.W., Parker, R.L., Wentworth, J., Madankumar, R., Saffer, C., Das, A.F., Craig, J.A., Chudova, D.I., Devers, P.L., and Jones, K.W. (2014). DNA sequencing versus standard prenatal aneuploidy screening. New England Journal of Medicine 370, 799-808.
92. Heim, S., and Mitelman, F. (2015). Cancer cytogenetics: chromosomal and molecular genetic aberrations of tumor cells.(John Wiley & Sons).
93. Kim, T.-M., Xi, R., Luquette, L.J., Park, R.W., Johnson, M.D., and Park, P.J. (2013). Functional genomic analysis of chromosomal aberrations in a compendium of 8000 cancer genomes. Genome research 23, 217-227.
94. Francis, G., and Stein, S. (2015). Circulating cell-free tumour DNA in the management of cancer. International journal of molecular sciences 16, 14122-14142.
95. Drets, M.E., and Shaw, M.W. (1971). Specific banding patterns of human chromosomes. Proceedings of the National Academy of Sciences 68, 2073-2077.
96. Landegent, J., in de Wal, N.J., van Omment, G.-J., Baas, F., de Vijlderi, J., Van Duijn, P., and Van der Ploeg, M. (1985). Chromosomal localization of a unique gene by non-autoradiographic in situ hybridization.
97. Schrock, E., Du Manoir, S., Veldman, T., and Schoell, B. (1996). Multicolor spectral karotyping of human chromosomes. Science 273, 494.
98. Speicher, M.R., Ballard, S.G., and Ward, D.C. (1996). Karyotyping human chromosomes by combinatorial multi-fluor FISH. Nature genetics 12, 368-375.
99. Pinkel, D., Segraves, R., Sudar, D., Clark, S., Poole, I., Kowbel, D., Collins, C., Kuo, W.-L., Chen, C., and Zhai, Y. (1998). High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. Nature genetics 20, 207-211.
100. Pirooznia, M., Goes, F.S., and Zandi, P.P. (2015). Whole-genome CNV analysis: advances in computational approaches. Frontiers in genetics 6.
101. Talevich, E., Shain, A.H., Botton, T., and Bastian, B.C. (2014). CNVkit: Copy number detection and visualization for targeted sequencing using off-target reads. BioRxiv, 010876.
102. Kuilman, T., Velds, A., Kemper, K., Ranzani, M., Bombardelli, L., Hoogstraat, M., Nevedomskaya, E., Xu, G., de Ruiter, J., and Lolkema, M.P. (2015). CopywriteR: DNA copy number detection from off-target sequence data. Genome biology 16, 49.
103. Frampton, G.M., Fichtenholtz, A., Otto, G.A., Wang, K., Downing, S.R., He, J., Schnall-Levin, M., White, J., Sanford, E.M., and An, P. (2013). Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing. Nature biotechnology 31, 1023-1031.
104. Wagle, N., Berger, M.F., Davis, M.J., Blumenstiel, B., DeFelice, M., Pochanard, P., Ducar, M., Van Hummelen, P., MacConaill, L.E., and Hahn, W.C. (2012). High-throughput detection of actionable genomic alterations in clinical tumor samples by targeted, massively parallel sequencing. Cancer discovery 2, 82-93.
105. Kinde, I., Papadopoulos, N., Kinzler, K.W., and Vogelstein, B. (2012). FAST-SeqS: a simple and efficient method for the detection of aneuploidy by massively parallel sequencing. PloS ONE 7, e41162.

106. Belic, J., Koch, M., Ulz, P., Auer, M., Gerhalter, T., Mohan, S., Fischereder, K., Petru, E., Bauernhofer, T., and Geigl, J.B. (2015). Rapid identification of plasma DNA samples with increased ctDNA levels by a modified FAST-SeqS approach. Clinical chemistry 61, 838-849.
107. Gale, D., Plagnol, V., Lawson, A., Pugh, M., Smalley, S., Howarth, K., Madi, M., Durham, B., Kumanduri, V., and Lo, K. (2016). Analytical performance and validation of an enhanced TAm-Seq circulating tumor DNA sequencing assay. In. (AACR.
108. Grasso, C., Butler, T., Rhodes, K., Quist, M., Neff, T.L., Moore, S., Tomlins, S.A., Reinig, E., Beadling, C., and Andersen, M. (2015). Assessing copy number alterations in targeted, amplicon-based next-generation sequencing data. The Journal of Molecular Diagnostics 17, 53-63.
109. Leary, R.J., Sausen, M., Kinde, I., Papadopoulos, N., Carpten, J.D., Craig, D., O'Shaughnessy, J., Kinzler, K.W., Parmigiani, G., and Vogelstein, B. (2012). Detection of chromosomal alterations in the circulation of cancer patients with whole-genome sequencing. Science translational medicine 4, 162ra154-162ra154.
110. Heitzer, E., Ulz, P., Belic, J., Gutschi, S., Quehenberger, F., Fischereder, K., Benezeder, T., Auer, M., Pischler, C., and Mannweiler, S. (2013). Tumor-associated copy number changes in the circulation of patients with prostate cancer identified through whole-genome sequencing. Genome medicine 5, 30.
111. Kinde, I., Wu, J., Papadopoulos, N., Kinzler, K.W., and Vogelstein, B. (2011). Detection and quantification of rare mutations with massively parallel sequencing. Proceedings of the National Academy of Sciences 108, 9530-9535.
112. Consortium, G.P. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature 491, 56-65.
113. Aird, D., Ross, M.G., Chen, W.-S., Danielsson, M., Fennell, T., Russ, C., Jaffe, D.B., Nusbaum, C., and Gnirke, A. (2011). Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome biology 12, R18.
114. Cortes, C., and Vapnik, V. (1995). Support-vector networks. Machine learning 20, 273-297.
115. Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., and Leisch, F. (2015). e1071: Misc Functions of the Department of Statistics (e1071), TU Wien, 2014. R package version, 1.6-3.
116. Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nature Methods 9, 357-359.
117. Straver, R., Sistermans, E.A., Holstege, H., Visser, A., Oudejans, C.B., and Reinders, M.J. (2014). WISECONDOR: detection of fetal aberrations from shallow sequencing maternal plasma based on a within-sample comparison scheme. Nucleic acids research 42, e31-e31.
118. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25, 1754-1760.
119. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The sequence alignment/map format and SAMtools. Bioinformatics 25, 2078-2079.
120. Center, B.I.T.G.D.A. (2016). Analysis-ready standardized TCGA data from Broad GDAC Firehose 2016_01_28 run.

121. Knouse, K.A., Davoli, T., Elledge, S.J., and Amon, A. (2017). Aneuploidy in Cancer: Seq-ing Answers to Old Questions.
122. Hoang, M.L., Chen, C.-H., Sidorenko, V.S., He, J., Dickman, K.G., Yun, B.H., Moriya, M., Niknafs, N., Douville, C., and Karchin, R. (2013). Mutational signature of aristolochic acid exposure as revealed by whole-exome sequencing. Science translational medicine 5, 197ra102-197ra102.

Curriculum Vitae

Christopher B. Douville

Mailing Address
116 West University Parkway #1326
Baltimore, MD 21210
(937) 219-9599 | cdouvill@gmail.com

Permanent Address
3741 Westwind Drive
Dayton, OH 45440

EDUCATION

Johns Hopkins University — Baltimore, MD
Ph.D. Biomedical Engineering — May 2017

University of Michigan — Ann Arbor, MI
B.S.E. Chemical Engineering (GPA 3.703) — April 2011

EXPERIENCE

Johns Hopkins University (Dr. Rachel Karchin) — **2011-present**
P.h.D Candidate

- Developed novel machine learning methods and statistical models to interpret genetic variation that resulted in 11 peer reviewed research papers, a platform research presentation (American Society for Human Genetics), and four publically available bioinformatics tools.
- Collaborated with a team of research scientists and professional software developers to design the Cancer Related Analysis Variant Toolkit (CRAVAT). A webserver that prioritizes and annotates genetic variation discovered in sequencing studies (over 6500 users in 6 continents available: http://www.cravat.us/CRAVAT/).
- Developed the Variant Effect Scoring Tool (VEST and VEST-indel) to prioritize genetic variation and determine the potential impact on human health (available: http://www.cravat.us/CRAVAT).
- Mentored 2 high school students and 5 undergraduate students.
- Taught 2 courses (Models and Simulations; Foundations in Computational Biology).
- Won a National Human Genome Research Institute (NHGRI F31) **Ruth Kirschstein Predoctoral Fellowship**.

University of Michigan (Dr.'s Stewart Wang and Grace Su) — **2008-2011**
Research Assistant

- Developed non-invasive models to diagnosis liver cirrhosis from CT images.
- Created human anatomical models for Toyota and Honda automotive safety modelling engineers.

- Collaborated with medical doctors, professional software developers, automotive engineers, and medical students.

Air Force Research Laboratory: Sensors Directorate **2006-08**
Research Intern/Wright Scholar

- Developed models to identify various geometries from incoming radar signals.
- Prepared geometric CAD models to validate radar prediction algorithms.

LEADERSHIP

- University of Michigan Varsity Swimming and Diving Team Captain
- Division 1 Varsity Swimming and Diving
- Student Athlete Advisory Council
- Biomedical Engineering Ph.D. Council

COMMUNITY SERVICE

My Sister's Place, Our Daily Bread, Ann Arbor Student School Reading Days, CS Mott Hospital Volunteering

MEMBERSHIPS IN PROFESSIONAL SOCIETIES

American Society for Human Genetics, American Association for Cancer Research, American Society for Clinical Oncology, International Society for Computational Biology

SKILLS

Python, R, Bioinformatics, DNA/RNA sequencing, Data Mining, Machine Learning, Statistics, SQL, MATLAB, MIMICS, Linux

Publications

1. Huhdanpaa H, **Douville C**, Baum K, Krishnamurthy VN, Holcombe S, Enchakalody B, et al. Development of a quantitative method for the diagnosis of cirrhosis. Scandinavian journal of gastroenterology. 2011;46(12):1468-77.
2. Carter H, **Douville C**, Stenson PD, Cooper DN, Karchin R. Identifying Mendelian disease genes with the variant effect scoring tool. BMC genomics. 2013;14(3):1.
3. Hoang ML, Chen C-H, Sidorenko VS, He J, Dickman KG, Yun BH, Moriya M, Niknafs N, **Douville C**, et al. Mutational signature of aristolochic acid exposure as revealed by whole-exome sequencing. Science translational medicine. 2013;5(197):197ra02-ra02.
4. **Douville C**, Carter H, Kim R, Niknafs N, Diekhans M, Stenson PD, et al. CRAVAT: cancer-related analysis of variants toolkit. Bioinformatics. 2013;29(5):647-8.
5. Huhdanpaa HT, Zhang P, Krishnamurthy VN, **Douville C**, Enchakolody B, Chou C, et al. Quantitative Detection of Cirrhosis: Towards the Development of Computer-Assisted Detection Method. Journal of digital imaging. 2014;27(5):601-9.
6. Sharma N, Sosnay PR, Ramalho AS, **Douville C**, Franca A, Gottschalk LB, et al. Experimental assessment of splicing variants using expression minigenes and comparison with in silico predictions. Human mutation. 2014;35(10):1249-59.

7. Chen YC, **Douville C**, Wang C, Niknafs N, Yeo G, Beleva-Guthrie V, et al. A probabilistic model to predict clinical phenotypic traits from genome sequencing. PLOS Comput Biol. 2014;10(9):e1003825.
8. Masica DL, Li S, **Douville C**, Manola J, Ferris RL, Burtness B, et al. Predicting survival in head and neck squamous cell carcinoma from TP53 mutation. Human genetics. 2015;134(5):497-507.
9. Rettig EM, Chung CH, Bishop JA, Howard JD, Sharma R, Li RJ, **Douville C**, et al. Cleaved NOTCH1 expression pattern in head and neck squamous cell carcinoma is associated with NOTCH1 mutation, HPV status, and high-risk features. Cancer Prevention Research. 2015;8(4):287-95.
10. Roberts NJ, Norris AL, Petersen GM, Bondy ML, Brand R, Gallinger S, Kurtz RC, Olson SH, Rustgi AK, Schwartz AG, Stoffel EM, Syngal S, Zogopoulous G, Ali SZ, Axilbund J, Chaffee KG, Chen YC, Cote ML, Childs EJ, **Douville C**, et al. Whole genome sequencing defines the genetic heterogeneity of familial pancreatic cancer. Cancer discovery. 2015:CD-15-0402.
11. Springer S, Wang Y, Dal Molin M, Masica DL, Jiao Y, Kinde I, Blackford A, Raman SP, Wolfgang CL, Tomita T, Niknafs N, **Douville C**, et al. A combination of molecular markers and clinical features improve the classification of pancreatic cysts. Gastroenterology. 2015;149(6):1501-10.
12. Turner TN, **Douville C**, Kim D, Stenson PD, Cooper DN, Chakravarti A, et al. Proteins linked to autosomal dominant and autosomal recessive disorders harbor characteristic rare missense mutation distribution patterns. Human Molecular Genetics. 2015:ddv309.
13. **Douville C**, Masica DL, Stenson PD, Cooper DN, Gygax DM, Kim R, et al. Assessing the Pathogenicity of Insertion and Deletion Variants with the Variant Effect Scoring Tool (VEST‐Indel). Human mutation. 2016;37(1):28-35.